

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problems Mailbox.**



Please type a plus sign (+) in this box



TECH CENTER 1600/29

PTO/SB (12-97)

Approved for use through 9/30/00. OMB 0651-0031
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Modified Form 1449/PTO INFORMATION DISCLOSURE STATEMENT BY APPLICANT (use as many sheets as necessary)	Application Number	09/718,321
	Filing Date	November 22, 2000
	First Named Inventor	Martin Leach
	Group Art Unit	Not Yet Assigned 1631
	Examiner Name	Not Yet Assigned L4
	Attorney Docket Number	15966-599 (CURA-99)

U.S. PATENT DOCUMENTS							
Exam Initials	Cite No.	U.S. Patent Document No.	Issue Date	Name of Assignee	Class	Sub Class	Filing Date If Appropriate
cm	A1	5,856,104	01/05/99	Affymetrix, Inc.	435	6	March 7, 1997

FOREIGN PATENT DOCUMENTS							
Exam Initials	Cite No.	Foreign Patent Document Office	Number	Name of Applicant	Date of Publication	Translation Yes	No
cm	B1	EP	0 717,113 A2	Affymax Technologies, N.V.	19-06-1966		X
	B2	WO	95/11995	Affymax Technologies, N.V.	4 May 1995		X
	B3	WO	93/22456	Trustees of Dartmouth College	11 November 1993		X
	B4	WO	98/14470	Genetics Institute, Inc.	9 April 1998		X
	B5	WO	98/20165	Whitehead Institute for Biomedical Research	14 May 1998		X
	B6	WO	98/56954	Affymetrix, Inc.	17 December 1998		X
	B7	WO	98/14466	Progentior, Inc.	9 April 1998		X
	B8	WO	97/19212	HP-Chemie Pelzer Research and Development Ltd.	29 Mai 1997		X

OTHER PRIOR ART - NON-PATENT LITERATURE DOCUMENTS		
Exam Initials	Cite No.	Name of Author, Title (when appropriate), Publication, Volume, Page(s), Date, Etc.
cm	C1	Abravaya, K., J. J. Carrino, et al. (1995). "Detection of point mutations with a modified ligase chain reaction (Gap- LCR)." <u>Nucleic Acids Res</u> 23(4): 675-82.
	C2	Adams, M. D., J. M. Kelley, et al. (1991). "Complementary DNA sequencing: expressed sequence tags and human genome project." <u>Science</u> 252(5013): 1651-6.
	C3	Barany, F. (1991). "Genetic disease detection and DNA amplification using cloned thermostable ligase." <u>Proc Natl Acad Sci U S A</u> 88(1): 189-93.
	C4	Cotton, R. G., N. R. Rodrigues, et al. (1988). "Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations." <u>Proc Natl Acad Sci U S A</u> 85(12): 4397-401.
	C5	Evans, W. E. and M. V. Relling (1999). "Pharmacogenomics: translating functional genomics into rational therapeutics." <u>Science</u> 286(5439): 487-91.
	C6	Faham, M. and D. R. Cox (1995). "A novel in vivo method to detect DNA sequence variation." <u>Genome Res</u> 5(5): 474-82.
	C7	Fischer, S. G. and L. S. Lerman (1983). "DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory." <u>Proc Natl Acad Sci U S A</u> 80(6): 1579-83.



RECEIVED

Page 2 of 2

AUG 01 2002

OTHER PRIOR ART - NON-PATENT LITERATURE DOCUMENTS			TECH CENTER 1600/2901
Exam Initials	Cite No.	Name of Author, Title (when appropriate), Publication, Volume, Page(s), Date, Etc.	
apl	C8	Gibbs, R. A., P. N. Nguyen, et al. (1989). "Detection of single DNA base differences by competitive oligonucleotide priming." <u>Nucleic Acids Res</u> 17(7): 2437-48.*	
	C9	Kren, B. T., B. Parashar, et al. (1999). "Correction of the UDP-glucuronosyltransferase gene defect in the gunn rat model of crigler-najjar syndrome type I with a chimeric oligonucleotide." <u>Proc Natl Acad Sci U S A</u> 96(18): 10349-54.	
	C10	Landegren, U., R. Kaiser, et al. (1988). "A ligase-mediated gene detection technique." <u>Science</u> 241(4869): 1077-80.	
	C11	Maskos, U. and E. M. Southern (1993). "A novel method for the parallel analysis of multiple mutations in multiple samples." <u>Nucleic Acids Res</u> 21(9): 2269-70.	
	C12	Myers, R. M., Z. Larin, et al. (1985). "Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes." <u>Science</u> 230(4731): 1242-6.	
	C13	Newton, C. R., A. Graham, et al. (1989). "Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS)." <u>Nucleic Acids Res</u> 17(7): 2503-16. (ABSTRACT)	
	C14	Nikiforov, T. T., R. B. Rehde, et al. (1994). "Genetic Bit Analysis: a solid phase method for typing single nucleotide polymorphisms." <u>Nucleic Acids Res</u> 22(20): 4167-75.	
	C15	Orita, M., Y. Suzuki, et al. (1989). "Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction." <u>Genomics</u> 5(4): 874-9.	
	C16	Orita, M., H. Iwahana, et al. (1989). "Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms." <u>Proc Natl Acad Sci U S A</u> 86(8): 2766-70.	
	C17	Orum, H., P. E. Nielsen, et al. (1993). "Single base pair mutation analysis by PNA directed PCR clamping." <u>Nucleic Acids Res</u> 21(23): 5332-6. (ABSTRACT)	
	C18	Rhodes, M., R. Straw, et al. (1998). "A high-resolution microsatellite map of the mouse genome." <u>Genome Res</u> 8(5): 531-42.	
	C19	Saiki, R. K., P. S. Walsh, et al. (1989). "Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes." <u>Proc Natl Acad Sci U S A</u> 86(16): 6230-4.	
	C20	Syvanen, A. C., K. Aalto-Setälä, et al. (1990). "A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E." <u>Genomics</u> 8(4): 684-92. (ABSTRACT)	
	C21	Taillon-Miller, P., Z. Gu, et al. (1998). "Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms [In Process Citation]." <u>Genome Res</u> 8(7): 748-54.	
	C22	Thiede, C., E. Bayerdorffer, et al. (1996). "Simple and sensitive detection of mutations in the ras proto-oncogenes using PNA-mediated PCR clamping." <u>Nucleic Acids Res</u> 24(5): 983-4.	
	C23	Wagner, R., P. Debbie, et al. (1995). "Mutation detection using immobilized mismatch binding protein (MutS)." <u>Nucleic Acids Res</u> 23(19): 3944-8.	
	C24	Wallace, R. B., J. Shaffer, et al. (1979). "Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch." <u>Nucleic Acids Res</u> 6(11): 3543-57.	
	C25	Youil, R., B. W. Kemper, et al. (1995). "Screening for mutations by enzyme mismatch cleavage with T4 endonuclease VII." <u>Proc Natl Acad Sci U S A</u> 92(1): 87-91.	

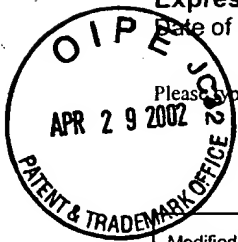
* a copy of this reference is not provided as it was previously cited by or submitted to the office in a prior application, Serial No. _____, filed _____, and relied upon for an earlier filing date under 35 U.S.C. §120 (continuation, continuation-in-part, and divisional applications).

Examiner Signature		Date Considered	9/25/07
--------------------	--	-----------------	---------

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; Draw line through citation if not in conformance and not considered.

Include copy of this form with next communication to applicant.

TRADOCs:1443156.1(%XJ_01!.DOC)



Express Mail No. EV058075585US

Date of Deposit: April 29, 2002

Please type a plus sign (+) in this box



RECEIVED

AUG 02 2002

TECH CENTER 1600/2900

RECEIVED
Page 1 of 1
MAY 03 2002
TECH CENTER 1600/2900
PTO 12-97
Approved for use through 9/30/00. OMB 0657-0031
Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Modified Form 1449/PTO INFORMATION DISCLOSURE STATEMENT BY APPLICANT (use as many sheets as necessary)	Application Number	09/718,321
	Filing Date	11/22/2000
	First Named Inventor	Shimkets
	Group Art Unit	4645 1631
	Examiner Name	Not Yet Assigned Ly
	Attorney Docket Number	15966-599

U.S. PATENT DOCUMENTS							
Exam Initials	Cite No.	U.S. Patent Document No.	Issue Date	Name of Patentee(s) or Applicant(s)	Class	Sub Class	Filing Date If Appropriate

FOREIGN PATENT DOCUMENTS							
Exam Initials	Cite No.	Foreign Patent Document Office Number	Name of Patentee(s) or Applicant(s)		Date of Publication	Translation Yes No	

OTHER PRIOR ART - NON PATENT LITERATURE DOCUMENTS		
Exam Initials	Cite No.	Name of Author, Title (when appropriate), Publication, Volume, Page(s), Date, Etc.
	C30	SWISS-PROT Database Accession Number: P55060 (10/01/96).
	C31	International Preliminary Examination Report for PCT/US00/32311. Mailed on April 9, 2002.

* a copy of this reference is not provided as it was previously cited by or submitted to the office in a prior application, Serial No. _____, filed _____, and relied upon for an earlier filing date under 35 U.S.C. §120 (continuation, continuation-in-part, and divisional applications).

Examiner Signature		Date Considered	9/25/03
---------------------------	--	------------------------	---------

EXAMINER: Initial if reference considered, whether or not citation is in conformance with MPEP 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to applicant.



US005856104A

United States Patent [19]

Chee et al.

[11] Patent Number: **5,856,104**[45] Date of Patent: **Jan. 5, 1999**[54] **POLYMORPHISMS IN THE GLUCOSE-6
PHOSPHATE DEHYDROGENASE LOCUS**[75] Inventors: **Mark Chee; Jian-Bing Fan**, both of
Palo Alto, Calif.[73] Assignee: **Affymetrix, Inc.**, Santa Clara, Calif.[21] Appl. No.: **813,508**[22] Filed: **Mar. 7, 1997****Related U.S. Application Data**

[60] Provisional application No. 60/029,374, Oct. 28, 1996.

[51] Int. Cl.⁶ **C12Q 1/68; C07H 21/04**[52] U.S. Cl. **435/6; 435/91.1; 435/91.2;
536/23.5; 536/24.31; 536/24.33**[58] Field of Search **435/6, 91.1, 91.2,
435/183; 536/23.5, 24.31, 24.33; 935/1,
8, 11, 76, 77**[56] **References Cited****U.S. PATENT DOCUMENTS**

5,206,137	4/1993	Ip et al.	435/6
5,292,639	3/1994	Beitz et al.	435/6
5,468,610	11/1995	Polymeropoulos et al.	435/6

OTHER PUBLICATIONS

Nierman and Maglott, eds., ATCC/NIH Repository Catalogue of Human and Mouse DNA Probes and Libraries, Eighth edition, pp. 1-70., 1989.

Sigma Molecular Biology catalog, p. 54., 1989.

Chen et al., "Long-range sequence analysis in human chromosome Xq28: thirteen known and six candidate genes in 219.4 kb of high GC DNA between the RCP/GCP and G6PD loci," *Hum. Mol. Genet.* vol. 5, no. 5, pp. 659-668; sequence comparison with Table 2 and, 1996.

Cappellini, Maria Domenica et al., "Multiple G6PD Mutations Are Associated With a Clinical and Biochemical Phenotype Similar to That of G6PD Mediterranean", *Blood*, vol. 87, (May 1, 1996) No. 9, pp. 3953-3958.

Fan, Jian-Bing et al., "Screening the Human Genome for Single-nucleotide Polymorphisms by Hybridization to High-density Oligonucleotide Arrays", *Mutation Detection '97*, 4th International Workshop, (May 29-Jun. 2, 1997) 1 page.

Kwok, Pui-Yan et al., "Increasing the Information Content of STS-Based Genome Maps: Identifying Polymorphisms in Mapped STSs", *Genomics*, (1996) vol. 31, Article No. 0019, pp. 123-126.

Primary Examiner—Bradley L. Sisson*Attorney, Agent, or Firm*—Townsend & Townsend & Crew[57] **ABSTRACT**

The invention provides nucleic acid segments of the glucose-6 phosphate dehydrogenase locus of the human genome including polymorphic sites. Allele-specific primers and probes hybridizing to regions flanking these sites are also provided. The nucleic acids, primers and probes are used in applications such as forensics, paternity testing, medicine and genetic analysis.

13 Claims, No Drawings

POLYMORPHISMS IN THE GLUCOSE-6 PHOSPHATE DEHYDROGENASE LOCUS

CROSS-REFERENCE TO RELATED APPLICATION

The present application claims priority from provisional application 60/029,374, filed Oct. 28, 1996, which is incorporated by reference in its entirety for all purposes.

BACKGROUND OF THE INVENTION

The genomes of all organisms undergo spontaneous mutation in the course of their continuing evolution generating variant forms of progenitor sequences (Gusella, *Ann. Rev. Biochem.* 55, 831-854 (1986)). The variant form may confer an evolutionary advantage or disadvantage relative to a progenitor form or may be neutral. In some instances, a variant form confers a lethal disadvantage and is not transmitted to subsequent generations of the organism. In other instances, a variant form confers an evolutionary advantage to the species and is eventually incorporated into the DNA of many or most members of the species and effectively becomes the progenitor form. In many instances, both progenitor and variant form(s) survive and co-exist in a species population. The coexistence of multiple forms of a sequence gives rise to polymorphisms.

Several different types of polymorphism have been reported. A restriction fragment length polymorphism (RFLP) means a variation in DNA sequence that alters the length of a restriction fragment as described in Botstein et al., *Am. J. Hum. Genet.* 32, 314-331 (1980). The restriction fragment length polymorphism may create or delete a restriction site, thus changing the length of the restriction fragment. RFLPs have been widely used in human and animal genetic analyses (see WO 90/13668; WO90/11369; Donis-Keller, *Cell* 51, 319-337 (1987); Lander et al., *Genetics* 121, 85-99 (1989)). When a heritable trait can be linked to a particular RFLP, the presence of the RFLP in an individual can be used to predict the likelihood that the animal will also exhibit the trait.

Other polymorphisms take the form of short tandem repeats (STRs) that include tandem di-, tri- and tetranucleotide repeated motifs. These tandem repeats are also referred to as variable number tandem repeat (VNTR) polymorphisms. VNTRs have been used in identity and paternity analysis (U.S. Pat. No. 5,075,217; Armour et al., *FEBS Lett.* 307, 113-115 (1992); Horn et al., WO 91/14003; Jeffreys, EP 370,719), and in a large number of genetic mapping studies.

Other polymorphisms take the form of single nucleotide variations between individuals of the same species. Such polymorphisms are far more frequent than RFLPs, STRs and VNTRs. Some single nucleotide polymorphisms occur in protein-coding sequences, in which case, one of the polymorphic forms may give rise to the expression of a defective or other variant protein and, potentially, a genetic disease. Examples of genes, in which polymorphisms within coding sequences give rise to genetic disease include β -globin (sickle cell anemia) and CFTR (cystic fibrosis). Other single nucleotide polymorphisms occur in noncoding regions. Some of these polymorphisms may also result in defective protein expression (e.g., as a result of defective splicing). Other single nucleotide polymorphisms have no phenotypic effects.

Single nucleotide polymorphisms can be used in the same manner as RFLPs, and VNTRs but offer several advantages. Single nucleotide polymorphisms occur with greater fre-

quency and are spaced more uniformly throughout the genome than other forms of polymorphism. The greater frequency and uniformity of single nucleotide polymorphisms means that there is a greater probability that such a polymorphism will be found in close proximity to a genetic locus of interest than would be the case for other polymorphisms. Also, the different forms of characterized single nucleotide polymorphisms are often easier to distinguish than other types of polymorphism (e.g., by use of assays employing allele-specific hybridization probes or primers).

Despite the increased amount of nucleotide sequence data being generated in recent years, only a minute proportion of the total repository of polymorphisms in humans and other organisms has so far been identified. The paucity of polymorphisms hitherto identified is due to the large amount of work required for their detection by conventional methods. For example, a conventional approach to identifying polymorphisms might be to sequence the same stretch of oligonucleotides in a population of individuals by didoxy sequencing. In this type of approach, the amount of work increases in proportion to both the length of sequence and the number of individuals in a population and becomes impractical for large stretches of DNA or large numbers of persons.

SUMMARY OF THE INVENTION

The invention provides nucleic acid segments of between 10 and 100 bases containing at least 10, 15 or 20 contiguous amino acids from any of the sequences shown in any of TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29) including a polymorphic site. Complements of these segments are also included. The segments can be DNA or RNA, and can be double- or single-stranded. Some segments are 10-20 or 10-50 bases long. Preferred segments include a diallelic polymorphic 25 site.

The invention further provides allele-specific oligonucleotides that hybridizes to a sequence shown in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29) or its complement. These oligonucleotides can be probes or primers.

The invention further provides a method of analyzing a nucleic acid from an individual. The method determines which base is present at any one of the polymorphic sites shown in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29). Optionally, a set of bases occupying a set of the polymorphic sites shown in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29).

11 (SEQ ID NOS:28 and 29). is determined. This type of analysis can be performed on a plurality of individuals who are tested for the presence of a disease phenotype. The presence or absence of disease phenotype can then be correlated with a base or set of bases present at the polymorphic sites in the individuals tested.

DEFINITIONS

An oligonucleotide can be DNA or RNA, and single- or double-stranded. Oligonucleotides can be naturally occurring or synthetic, but are typically prepared by synthetic means. Preferred oligonucleotides of the invention include segments of DNA, or their complements including any one of the polymorphic sites shown in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29). The segments are usually between 5 and 100 bases, and often between 5-10, 5-20, 10-20, 10-50, 20-50 or 20-100 bases. The polymorphic site can occur within any position of the segment. The segments can be from any of the allelic forms of DNA shown in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29).

Hybridization probes are oligonucleotides capable of binding in a base-specific manner to a complementary strand of nucleic acid. Such probes include peptide nucleic acids, as described in Nielsen et al., *Science* 254, 1497-1500 (1991).

The term primer refers to a single-stranded oligonucleotide capable of acting as a point of initiation of template-directed DNA synthesis under appropriate conditions (i.e., in the presence of four different nucleoside triphosphates and an agent for polymerization, such as, DNA or RNA polymerase or reverse transcriptase) in an appropriate buffer and at a suitable temperature. The appropriate length of a primer depends on the intended use of the primer but typically ranges from 15 to 30 nucleotides. Short primer molecules generally require cooler temperatures to form sufficiently stable hybrid complexes with the template. A primer need not reflect the exact sequence of the template but must be sufficiently complementary to hybridize with a template. The term primer site refers to the area of the target DNA to which a primer hybridizes. The term primer pair means a set of primers including a 5' upstream primer that hybridizes with the 5' end of the DNA sequence to be amplified and a 3', downstream primer that hybridizes with the complement of the 3' end of the sequence to be amplified.

Linkage describes the tendency of genes, alleles, loci or genetic markers to be inherited together as a result of their location on the same chromosome, and can be measured by percent recombination between the two genes, alleles, loci or genetic markers.

Polymorphism refers to the occurrence of two or more genetically determined alternative sequences or alleles in a population. A polymorphic marker or site is the locus at which divergence occurs. Preferred markers have at least two alleles, each occurring at frequency of greater than 1%, and more preferably greater than 10% or 20% of a selected

population. A polymorphic locus may be as small as one base pair. Polymorphic markers include restriction fragment length polymorphisms, variable number of tandem repeats (VNTR's), hypervariable regions, minisatellites, dinucleotide repeats, trinucleotide repeats, tetranucleotide repeats, simple sequence repeats, and insertion elements such as Alu. The first identified allelic form is arbitrarily designated as a the reference form and other allelic forms are designated as alternative or variant alleles. The allelic form occurring most frequently in a selected population is sometimes referred to as the wildtype form. Diploid organisms may be homozygous or heterozygous for allelic forms. A diallelic polymorphism has two forms. A triallelic polymorphism has three forms.

A single nucleotide polymorphism occurs at a polymorphic site occupied by a single nucleotide, which is the site of variation between allelic sequences. The site is usually preceded by and followed by highly conserved sequences of the allele (e.g., sequences that vary in less than 1/100 or 1/1000 members of the populations).

A single nucleotide polymorphism usually arises due to substitution of one nucleotide for another at the polymorphic site. A transition is the replacement of one purine by another purine or one pyrimidine by another pyrimidine. A transversion is the replacement of a purine by a pyrimidine or vice versa. Single nucleotide polymorphisms can also arise from a deletion of a nucleotide or an insertion of a nucleotide relative to a reference allele.

Hybridizations are usually performed under stringent conditions, for example, at a salt concentration of no more than 1M and a temperature of at least 25° C. For example, conditions of 5X SSPE (750 mM NaCl, 50 mM NaPhosphate, 5 mM EDTA, pH 7.4) and a temperature of 25°-30° C. are suitable for allele-specific probe hybridizations.

An isolated nucleic acid means an object species (invention) that is the predominant species present (i.e., on a molar basis it is more abundant than any other individual species in the composition). Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90 percent (on a molar basis) of all macromolecular species present. Most preferably, the object species is purified to essential homogeneity (contaminant species cannot be detected in the composition by conventional detection methods).

DESCRIPTION OF THE PRESENT INVENTION

I. Novel Polymorphisms of the Invention

The human glucose-6-phosphate dehydrogenase locus (G6PD) encompasses more than 50,000 bp and resides on the X chromosome. A complete prototypical sequence of the G6PD locus has been published. That locus has remained relatively unexplored due to the cost and difficulty of conventional sequence analysis. The published sequence shows that the G6PD locus contains at least two genes, the G6PD gene and the 2₁₉ gene. Those genes span approximately 16,000 bp and 10,000 bp, respectively. The enzyme G6PD play a fundamental role in glucose metabolism. The function of the 2-19 polypeptide product, however, has not been shown.

The present application provides 10 polymorphisms at 10 sequence tagged sites in the human G6PD locus. Table 1 shows the base occupied at those ten sites in 10 individuals. The sequences flanking each of these polymorphisms are shown in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17),

TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29). The polymorphic site is flanked by bold lines in the table. The sequences designated M1-M10 represent novel allelic variants of this site. The designation N as it appears in Tables 1-11 means the identity of a base was not determined.

II. Analysis of Polymorphisms

A. Preparation of Samples

Polymorphisms are detected in a target nucleic acid from an individual being analyzed. For assay of genomic DNA, virtually any biological sample (other than pure red blood cells) is suitable. For example, convenient tissue samples include whole blood, semen, saliva, tears, urine, fecal material, sweat, buccal, skin and hair. For assay of cDNA or mRNA, the tissue sample must be obtained from an organ in which the target nucleic acid is expressed.

Many of the methods described below require amplification of DNA from target samples. This can be accomplished by e.g., PCR. See generally *PCR Technology: Principles and Applications for DNA Amplification* (ed. H. A. Erlich, Freeman Press, N.Y., N.Y., 1992); *PCR Protocols: A Guide to Methods and Applications* (eds. Innis, et al., Academic Press, San Diego, Calif., 1990); Mattila et al., *Nucleic Acids Res.* 19, 4967 (1991); Eckert et al., *PCR Methods and Applications* 1, 17 (1991); *PCR* (eds. McPherson et al., IRL Press, Oxford); and U.S. Pat. No. 4,683,202 (each of which is incorporated by reference for all purposes).

Other suitable amplification methods include the ligase chain reaction (LCR) (see Wu and Wallace, *Genomics* 4, 560 (1989), Landegren et al., *Science* 241, 1077 (1988), transcription amplification (Kwoh et al., *Proc. Natl. Acad. Sci. USA* 86, 1173 (1989)), and self-sustained sequence replication (Guatelli et al., *Proc. Nat. Acad. Sci. USA*, 87, 1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

B. Detection of Polymorphisms in Target DNA

There are two distinct types of analysis depending whether a polymorphism in question has already been characterized. The first type of analysis is sometimes referred to as de novo characterization. This analysis compares target sequences in different individuals to identify points of variation, i.e., polymorphic sites. By analyzing a groups of individuals representing the greatest ethnic diversity among humans and greatest breed and species variety in plants and animals, patterns characteristic of the most common alleles/haplotypes of the locus can be identified, and the frequencies of such populations in the population determined. Additional allelic frequencies can be determined for subpopulations characterized by criteria such as geography, race, or gender. The de novo identification of the polymorphisms of the invention is described in the Examples section. The second type of analysis is determining which form(s) of a characterized polymorphism are present in individuals under test. There are a variety of suitable procedures, which are discussed in turn.

1. Allele-Specific Probes

The design and use of allele-specific probes for analyzing polymorphisms is described by e.g., Saiki et al., *Nature* 324, 163-166 (1986); Dattagupta, EP 235,726, Saiki, WO 89/11548. Allele-specific probes can be designed that hybridize to a segment of target DNA from one individual

but do not hybridize to the corresponding segment from another individual due to the presence of different polymorphic forms in the respective segments from the two individuals. Hybridization conditions should be sufficiently stringent that there is a significant difference in hybridization intensity between alleles, and preferably an essentially binary response, whereby a probe hybridizes to only one of the alleles. Some probes are designed to hybridize to a segment of target DNA such that the polymorphic site aligns with a central position (e.g., in a 15 mer at the 7 position; in a 16 mer, at either the 8 or 9 position) of the probe. This design of probe achieves good discrimination in hybridization between different allelic forms.

Allele-specific probes are often used in pairs, one member of a pair showing a perfect match to a reference form of a target sequence and the other member showing a perfect match to a variant form. Several pairs of probes can then be immobilized on the same support for simultaneous analysis of multiple polymorphisms within the same target sequence.

2. Tiling Arrays

The polymorphisms can also be identified by hybridization to nucleic acid arrays, some example of which are described by WO 95/11995 (incorporated by reference in its entirety for all purposes). One form of such arrays is described in the Examples section in connection with de novo identification of polymorphisms. The same array or a different array can be used for analysis of characterized polymorphisms. WO 95/11995 also describes subarrays that are optimized for detection of a variant forms of a precharacterized polymorphism. Such a subarray contains probes designed to be complementary to a second reference sequence, which is an allelic variant of the first reference sequence. The second group of probes is designed by the same principles as described in the Examples except that the probes exhibit complementarity to the second reference sequence. The inclusion of a second group (or further groups) can be particularly useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (i.e., two or more mutations within 9 to 21 bases).

3. Allele-Specific Primers

An allele-specific primer hybridizes to a site on target DNA overlapping a polymorphism and only primes amplification of an allelic form to which the primer exhibits perfect complementarity. See Gibbs, *Nucleic Acid Res.* 17, 2427-2448 (1989). This primer is used in conjunction with a second primer which hybridizes at a distal site. Amplification proceeds from the two primers leading to a detectable product signifying the particular allelic form is present. A control is usually performed with a second pair of primers, one of which shows a single base mismatch at the polymorphic site and the other of which exhibits perfect complementarity to a distal site. The single-base mismatch prevents amplification and no detectable product is formed. The method works best when the mismatch is included in the 3'-most position of the oligonucleotide aligned with the polymorphism because this position is most destabilizing to elongation from the primer. See, e.g., WO 93/22456.

4. Direct-Sequencing

The direct analysis of the sequence of polymorphisms of the present invention can be accomplished using either the dideoxy chain termination method or the Maxam Gilbert method (see Sambrook et al., *Molecular Cloning, A Laboratory Manual* (2nd Ed., CSHP, New York 1989); Zyskind et al., *Recombinant DNA Laboratory Manual*, (Acad. Press, 1988)).

5. Denaturing Gradient Gel Electrophoresis

Amplification products generated using the polymerase chain reaction can be analyzed by the use of denaturing gradient gel electrophoresis. Different alleles can be identified based on the different sequence-dependent melting properties and electrophoretic migration of DNA in solution. Erlich, ed., *PCR Technology, Principles and Applications for DNA Amplification*, (W.H. Freeman and Co, New York, 1992), Chapter 7.

6. Single-Strand Conformation Polymorphism Analysis

Alleles of target sequences can be differentiated using single-strand conformation polymorphism analysis, which identifies base differences by alteration in electrophoretic migration of single stranded PCR products, as described in Orita et al., *Proc. Nat. Acad. Sci.* 86, 2766-2770 (1989). Amplified PCR products can be generated as described above, and heated or otherwise denatured, to form single stranded amplification products. Single-stranded nucleic acids may refold or form secondary structures which are partially dependent on the base sequence. The different electrophoretic mobilities of single-stranded amplification products can be related to base-sequence difference between alleles of target sequences.

III. Methods of Use

After determining polymorphic form(s) present in an individual at one or more polymorphic sites, this information can be used in a number of methods.

A. Forensics

Determination of which polymorphic forms occupy a set of polymorphic sites in an individual identifies a set of polymorphic forms that distinguishes the individual. See generally National Research Council, *The Evaluation of Forensic DNA Evidence* (Eds. Pollard et al., National Academy Press, DC, 1996). Since the polymorphic sites are within a 50,000 bp region in the human genome, the probability of recombination between these polymorphic sites is low. That low probability means the haplotype (the set of all 10 polymorphic sites) set forth in this application should be inherited without change for at least several generations. The more sites that are analyzed the lower the probability that the set of polymorphic forms in one individual is the same as that in an unrelated individual. Preferably, if multiple sites are analyzed, the sites are unlinked. Thus, polymorphisms of the invention are often used in conjunction with polymorphisms in distal genes. Preferred polymorphisms for use in forensics are diallelic because the population frequencies of two polymorphic forms can usually be determined with greater accuracy than those of multiple polymorphic forms at multi-allelic loci.

The capacity to identify a distinguishing or unique set of forensic markers in an individual is useful for forensic analysis. For example, one can determine whether a blood sample from a suspect matches a blood or other tissue sample from a crime scene by determining whether the set of polymorphic forms occupying selected polymorphic sites is the same in the suspect and the sample. If the set of polymorphic markers does not match between a suspect and a sample, it can be concluded (barring experimental error) that the suspect was not the source of the sample. If the set of markers does match, one can conclude that the DNA from the suspect is consistent with that found at the crime scene. If frequencies of the polymorphic forms at the loci tested have been determined (e.g., by analysis of a suitable population of individuals), one can perform a statistical analysis to determine the probability that a match of suspect and crime scene sample would occur by chance.

$p(ID)$ is the probability that two random individuals have the same polymorphic or allelic form at a given polymorphic

site. In diallelic loci, four genotypes are possible: AA, AB, BA, and BB. If alleles A and B occur in a haploid genome of the organism with frequencies x and y , the probability of each genotype in a diploid organism are (see WO 95/12607):

$$\text{Homozygote: } p(AA)=x^2$$

$$\text{Homozygote: } p(BB)=y^2=(1-x)^2$$

$$\text{Single Heterozygote: } p(AB)=p(BA)=xy=x(1-x)$$

$$\text{Both Heterozygotes: } p(AB+BA)=2xy=2x(1-x)$$

The probability of identity at one locus (i.e., the probability that two individuals, picked at random from a population will have identical polymorphic forms at a given locus) is given by the equation:

$$p(ID)=(x^2)^2+(2xy)^2+(y^2)^2.$$

These calculations can be extended for any number of polymorphic forms at a given locus. For example, the probability of identity $p(ID)$ for a 3-allele system where the alleles have the frequencies in the population of x , y and z , respectively, is equal to the sum of the squares of the genotype frequencies:

$$p(ID)=x^4+(2xy)^2+(2yz)^2+(2xz)^2+z^4+y^4$$

In a locus of n alleles, the appropriate binomial expansion is used to calculate $p(ID)$ and $p(exc)$.

The cumulative probability of identity (cum $p(ID)$) for each of multiple unlinked loci is determined by multiplying the probabilities provided by each locus.

$$\text{cum } p(ID)=p(ID1)p(ID2)p(ID3) \dots p(IDn)$$

The cumulative probability of non-identity for n loci (i.e., the probability that two random individuals will be different at 1 or more loci) is given by the equation:

$$\text{cum } p(\text{nonID})=1-\text{cum } p(ID).$$

If several polymorphic loci are tested, the cumulative probability of non-identity for random individuals becomes very high (e.g., one billion to one). Such probabilities can be taken into account together with other evidence in determining the guilt or innocence of the suspect.

B. Paternity Testing

The object of paternity testing is usually to determine whether a male is the father of a child. In most cases, the mother of the child is known and thus, the mother's contribution to the child's genotype can be traced. Paternity testing investigates whether the part of the child's genotype not attributable to the mother is consistent with that of the putative father. Paternity testing can be performed by analyzing sets of polymorphisms in the putative father and the child.

If the set of polymorphisms in the child attributable to the father does not match the putative father, it can be concluded, barring experimental error, that the putative father is not the real father. If the set of polymorphisms in the child attributable to the father does match the set of polymorphisms of the putative father, a statistical calculation can be performed to determine the probability of coincidental match.

The probability of parentage exclusion (representing the probability that a random male will have a polymorphic form at a given polymorphic site that makes him incompatible as the father) is given by the equation (see WO 95/12607):

$$p(exc)=xy(1-xy)$$

where x and y are the population frequencies of alleles A and B of a diallelic polymorphic site.

(At a triallelic site $p(exc)=xy(1-xy)+yz(1-yz)+xz(1-xz)+3xyz(1-xyz)$), where x , y and z are the respective population frequencies of alleles A, B and C).

The probability of non-exclusion is

$$p(non-exc)=1-p(exc)$$

The cumulative probability of non-exclusion (representing the value obtained when n loci are used) is thus:

$$cum\ p(non-exc)=p(non-exc1)p(non-exc2)p(non-exc3)\dots p(non-exc n)$$

The cumulative probability of exclusion for n loci (representing the probability that a random male will be excluded)

$$cum\ p(exc)=1-cum\ p(non-exc).$$

If several polymorphic loci are included in the analysis, the cumulative probability of exclusion of a random male is very high. This probability can be taken into account in assessing the liability of a putative father whose polymorphic marker set matches the child's polymorphic marker set attributable to his/her father.

C. Correlation of Polymorphisms with Phenotypic Traits

The polymorphisms of the invention may contribute to the phenotype of an organism in different ways. Some polymorphisms occur within a protein coding sequence and contribute to phenotype by affecting protein structure. The effect may be neutral, beneficial or detrimental, or both beneficial and detrimental, depending on the circumstances. For example, a heterozygous sickle cell mutation confers resistance to malaria, but a homozygous sickle cell mutation is usually lethal. Other polymorphisms occur in noncoding regions but may exert phenotypic effects indirectly via influence on replication, transcription, and translation. A single polymorphism may affect more than one phenotypic trait. Likewise, a single phenotypic trait may be affected by polymorphisms in different genes. Further, some polymorphisms predispose an individual to a distinct mutation that is causally related to a certain phenotype.

Phenotypic traits include diseases that have known but hitherto unmapped genetic components. Phenotypic traits also include symptoms of, or susceptibility to, multifactorial diseases of which a component is or may be genetic, such as autoimmune diseases, inflammation, cancer, diseases of the nervous system, and infection by pathogenic microorganisms. Some examples of autoimmune diseases include rheumatoid arthritis, multiple sclerosis, diabetes (insulin-dependent and non-independent), systemic lupus erythematosus and Graves disease. Some examples of cancers include cancers of the bladder, brain, breast, colon, esophagus, kidney, leukemia, liver, lung, oral cavity, ovary,

pancreas, prostate, skin, stomach and uterus. Phenotypic traits also include characteristics such as longevity, appearance (e.g., baldness, obesity), strength, speed, endurance, fertility, and susceptibility or receptivity to particular drugs or therapeutic treatments.

Correlation is performed for a population of individuals who have been tested for the presence or absence of a phenotypic trait of interest and for polymorphic markers sets. To perform such analysis, the presence or absence of a set of polymorphisms (i.e. a polymorphic set) is determined for a set of the individuals, some of whom exhibit a particular trait, and some of which exhibit lack of the trait. The alleles of each polymorphism of the set are then reviewed to determine whether the presence or absence of a particular allele is associated with the trait of interest. Correlation can be performed by standard statistical methods such as a χ^2 -squared test and statistically significant correlations between polymorphic form(s) and phenotypic characteristics are noted. For example, it might be found that the presence of allele A1 at polymorphism A correlates with heart disease. As a further example, it might be found that the combined presence of allele A1 at polymorphism A and allele B1 at polymorphism B correlates with increased milk production of a farm animal.

Such correlations can be exploited in several ways. In the case of a strong correlation between a set of one or more polymorphic forms and a disease for which treatment is available, detection of the polymorphic form set in a human or animal patient may justify immediate administration of treatment, or at least the institution of regular monitoring of the patient. Detection of a polymorphic form correlated with serious disease in a couple contemplating a family may also be valuable to the couple in their reproductive decisions. For example, the female partner might elect to undergo in vitro fertilization to avoid the possibility of transmitting such a polymorphism from her husband to her offspring. In the case of a weaker, but still statistically significant correlation between a polymorphic set and human disease, immediate therapeutic intervention or monitoring may not be justified. Nevertheless, the patient can be motivated to begin simple life-style changes (e.g., diet, exercise) that can be accomplished at little cost to the patient but confer potential benefits in reducing the risk of conditions to which the patient may have increased susceptibility by virtue of variant alleles. Identification of a polymorphic set in a patient correlated with enhanced receptiveness to one of several treatment regimes for a disease indicates that this treatment regime should be followed.

For animals and plants, correlations between characteristics and phenotype are useful for breeding for desired characteristics. For example, Beitz et al., U.S. Pat. No. 5,292,639 discuss use of bovine mitochondrial polymorphisms in a breeding program to improve milk production in cows. To evaluate the effect of mtDNA D-loop sequence polymorphism on milk production, each cow was assigned a value of 1 if variant or 0 if wildtype with respect to a prototypical mitochondrial DNA sequence at each of 17 locations considered. Each production trait was analyzed individually with the following animal model:

$$Y_{ijklm}=\mu+YS_i+P_j+X_k+\beta_1+\dots+\beta_{17}+PE_n+a_n+e_p$$

where Y_{ijklm} is the milk, fat, fat percentage, SNF, SNF percentage, energy concentration, or lactation energy record; μ is an overall mean; YS_i is the effect common to all cows calving in year-season; X_k is the effect common to cows in either the high or average selection line; β_1 to β_{17} are the

binomial regressions of production record on mtDNA D-loop sequence polymorphisms; PE_n is permanent environmental effect common to all records of cow n ; a_n is effect of animal n and is composed of the additive genetic contribution of sire and dam breeding values and a Mendelian sampling effect; and e_p is a random residual. It was found that eleven of seventeen polymorphisms tested influenced at least one production trait. Bovines having the best polymorphic forms for milk production at these eleven loci are used as parents for breeding the next generation of the herd.

D. Genetic Mapping of Phenotypic Traits

The previous section concerns identifying correlations between phenotypic traits and polymorphisms that directly or indirectly contribute to those traits. The present section describes identification of a physical linkage between a genetic locus associated with a trait of interest and polymorphic markers that are not associated with the trait, but are in physical proximity with the genetic locus responsible for the trait and co-segregate with it. Such analysis is useful for mapping a genetic locus associated with a phenotypic trait to a chromosomal position, and thereby cloning gene(s) responsible for the trait. See Lander et al., *Proc. Natl. Acad. Sci. (USA)* 83, 7353-7357 (1986); Lander et al., *Proc. Natl. Acad. Sci. (USA)* 84, 2363-2367 (1987); Donis-Keller et al., *Cell* 51, 319-337 (1987); Lander et al., *Genetics* 121, 185-199 (1989). Genes localized by linkage can be cloned by a process known as directional cloning. See Wainwright, *Med. J. Australia* 159, 170-174 (1993); Collins, *Nature Genetics* 1, 3-6 (1992) (each of which is incorporated by reference in its entirety for all purposes).

Linkage studies are typically performed on members of a family. Available members of the family are characterized for the presence or absence of a phenotypic trait and for a set of polymorphic markers. The distribution of polymorphic markers in an informative meiosis is then analyzed to determine which polymorphic markers co-segregate with a phenotypic trait. See, e.g., Kerem et al., *Science* 245, 1073-1080 (1989); Monaco et al., *Nature* 316, 842 (1985); Yamoka et al., *Neurology* 40, 222-226 (1990); Rossiter et al., *FASEB Journal* 5, 21-27 (1991).

Linkage is analyzed by calculation of LOD (log of the odds) values. A lod value is the relative likelihood of obtaining observed segregation data for a marker and a genetic locus when the two are located at a recombination fraction θ , versus the situation in which the two are not linked, and thus segregating independently (Thompson & Thompson, *Genetics in Medicine* (5th ed, W.B. Saunders Company, Philadelphia, 1991); Strachan, "Mapping the human genome" in *The Human Genome* (BIOS Scientific Publishers Ltd, Oxford), Chapter 4). A series of likelihood ratios are calculated at various recombination fractions (θ), ranging from $\theta=0.0$ (coincident loci) to $\theta=0.50$ (unlinked). Thus, the likelihood at a given value of θ is: probability of data if loci linked at θ to probability of data if loci unlinked. The computed likelihoods are usually expressed as the \log_{10} of this ratio (i.e., a lod score). For example, a lod score of 3 indicates 1000:1 odds against an apparent observed linkage being a coincidence. The use of logarithms allows data collected from different families to be combined by simple addition. Computer programs are available for the calculation of lod scores for differing values of θ (e.g., LIPED, MLINK (Lathrop, *Proc. Nat. Acad. Sci. (USA)* 81, 3443-3446 (1984)). For any particular lod score, a recombination fraction may be determined from mathematical tables. See Smith et al., *Mathematical tables for research workers in human genetics* (Churchill, London, 1961); Smith, *Ann. Hum. Genet.* 32, 127-150 (1968). The value of

θ at which the lod score is the highest is considered to be the best estimate of the recombination fraction.

Positive lod score values suggest that the two loci are linked, whereas negative values suggest that linkage is less likely (at that value of θ) than the possibility that the two loci are unlinked. By convention, a combined lod score of +3 or greater (equivalent to greater than 1000:1 odds in favor of linkage) is considered definitive evidence that two loci are linked. Similarly, by convention, a negative lod score of -2 or less is taken as definitive evidence against linkage of the two loci being compared. Negative linkage data are useful in excluding a chromosome or a segment thereof from consideration. The search focuses on the remaining non-excluded chromosomal locations.

IV. Modified Polypeptides and Gene Sequences

The invention further provides variant forms of nucleic acids and corresponding proteins. The nucleic acids comprise at least ten contiguous bases of one of the sequences described in TABLE 2 (SEQ ID NOS:1 and 2), TABLE 3 (SEQ ID NOS:3 and 4), TABLE 4 (SEQ ID NOS:5 and 6), TABLE 5 (SEQ ID NOS:7-15), TABLE 6 (SEQ ID NOS:16 and 17), TABLE 7 (SEQ ID NOS:18 and 19), TABLE 8 (SEQ ID NOS:20 and 21), TABLE 9 (SEQ ID NOS:22-24), TABLE 10 (SEQ ID NOS:25-27) and TABLE 11 (SEQ ID NOS:28 and 29), designated M1-M10. Some nucleic acid encode full-length variant forms of proteins. Similarly, variant proteins have the prototypical amino acid sequences of encoded by nucleic acid sequence shown in Tables 2-11, designated M1-M10 (read so as to be in-frame with the full-length coding sequence of which it is a component).

Variant genes can be expressed in an expression vector in which a variant gene is operably linked to a native or other promoter. Usually, the promoter is a eukaryotic promoter for expression in a mammalian cell. The transcription regulation sequences typically include a heterologous promoter and optionally an enhancer which is recognized by the host. The selection of an appropriate promoter, for example trp, lac, phage promoters, glycolytic enzyme promoters and tRNA promoters, depends on the host selected. Commercially available expression vectors can be used. Vectors can include host-recognized replication systems, amplifiable genes, selectable markers, host sequences useful for insertion into the host genome, and the like.

The means of introducing the expression construct into a host cell varies depending upon the particular construction and the target host. Suitable means include fusion, conjugation, transfection, transduction, electroporation or injection, as described in Sambrook, supra. A wide variety of host cells can be employed for expression of the variant gene, both prokaryotic and eukaryotic. Suitable host cells include bacteria such as *E. coli*, yeast, filamentous fungi, insect cells, mammalian cells, typically immortalized, e.g., mouse, CHO, human and monkey cell lines and derivatives thereof. Preferred host cells are able to process the variant gene product to produce an appropriate mature polypeptide. Processing includes glycosylation, ubiquitination, disulfide bond formation, general post-translational modification, and the like.

The protein may be isolated by conventional means of protein biochemistry and purification to obtain a substantially pure product, i.e., 80, 95 or 99% free of cell component contaminants, as described in Jacoby, *Methods in Enzymology* Volume 104, Academic Press, New York (1984); Scopes, *Protein Purification, Principles and Practice*, 2nd Edition, Springer-Verlag, New York (1987); and Deutscher (ed), *Guide to Protein Purification, Methods in Enzymology*, Vol. 182 (1990). If the protein is secreted, it can be isolated

from the supernatant in which the host cell is grown. If not secreted, the protein can be isolated from a lysate of the host cells.

The invention further provides transgenic nonhuman animals capable of expressing an exogenous variant gene and/or having one or both alleles of an endogenous variant gene inactivated. Expression of an exogenous variant gene is usually achieved by operably linking the gene to a promoter and optionally an enhancer, and microinjecting the construct into a zygote. See Hogan et al., "Manipulating the Mouse Embryo, A Laboratory Manual," Cold Spring Harbor Laboratory. Inactivation of endogenous variant genes can be achieved by forming a transgene in which a cloned variant gene is inactivated by insertion of a positive selection marker. See Capecchi, *Science* 244, 1288-1292 (1989). The transgene is then introduced into an embryonic stem cell, where it undergoes homologous recombination with an endogenous variant gene. Mice and other rodents are preferred animals. Such animals provide useful drug screening systems.

In addition to substantially full-length polypeptides expressed by variant genes, the present invention includes biologically active fragments of the polypeptides, or analogs thereof, including organic molecules which simulate the interactions of the peptides. Biologically active fragments include any portion of the full-length polypeptide which confers a biological function on the variant gene product, including ligand binding, and antibody binding. Ligand binding includes binding by nucleic acids, proteins or polypeptides, small biologically active molecules, or large cellular structures.

Polyclonal and/or monoclonal antibodies that specifically bind to variant gene products but not to corresponding prototypical gene products are also provided. Antibodies can be made by injecting mice or other animals with the variant gene product or synthetic peptide fragments thereof. Monoclonal antibodies are screened as are described, for example, in Harlow & Lane, *Antibodies, A Laboratory Manual*, Cold Spring Harbor Press, New York (1988); Goding, *Monoclonal antibodies, Principles and Practice* (2d ed.) Academic Press, New York (1986). Monoclonal antibodies are tested for specific immunoreactivity with a variant gene product and lack of immunoreactivity to the corresponding prototypical gene product. These antibodies are useful in diagnostic assays for detection of the variant form, or as an active ingredient in a pharmaceutical composition.

V. Kits

The invention further provides kits comprising at least one allele-specific oligonucleotide as described above. Often, the kits contain one or more pairs of allele-specific oligonucleotides hybridizing to different forms of a polymorphism. In some kits, the allele-specific oligonucleotides are provided immobilized to a substrate. For example, the same substrate can comprise allele-specific oligonucleotide probes for detecting at least 10, 100 or all of the polymorphisms shown in Tables 2-11. Optional additional components of the kit include, for example, restriction enzymes, reverse-transcriptase or polymerase, the substrate nucleoside triphosphates, means used to label (for example, an avidinenzyme conjugate and enzyme substrate and chromogen if the label is biotin), and the appropriate buffers for reverse transcription, PCR, or hybridization reactions. Usually, the kit also contains instructions for carrying out the methods.

EXAMPLES

The polymorphisms set forth in this application were identified by hybridization to tiling arrays. Tiling arrays are described in PCT/US94/12305 (incorporated by reference in its entirety for all purposes). Tiling generally means the synthesis of a defined set of oligonucleotide probes that is made up of a sequence complementary to the sequence to be analyzed (the "target sequence"), as well as preselected variations of that sequence. The variations usually include substitution at one or more base positions with one or more nucleotides. Tiling strategies are discussed in WO 95/11995 (incorporated by reference in its entirety for all purposes). With a tiled array containing 4L probes one can query every position in a nucleotide containing L number of bases. A 4L tiled array, for example, contains L number of sets of 4 probes, i.e. 4L probes. Each set of 4 probes contains the perfect complement to a portion of the target sequence with a single substitution for each nucleotide at the same position in the probe. See also Chee, M., et. al., *Science*, October, 1996.

To detect the novel sequence tagged polymorphic sites provided in this application, we designed a P^{25,13} (25-mer probes having the interrogation position at base 13) 4L tiling array for the G6PD locus. Because the G6PD locus contains a large number of Alu sequences (repeat sequences), we simplified the tiled probe array by not probing the repetitive Alu sequences. To generate target sequence fragments, blood was collected from 10 individuals. Long range PCR amplification was carried out on genomic DNA. The amplicons were labeled, fragmented, and used to determine hybridization to the array.

All publications and patent applications cited above are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication or patent application were specifically and individually indicated to be so incorporated by reference. Although the present invention has been described in some detail by way of illustration and example for purposes of clarity and understanding, it will be apparent that certain changes and modifications may be practiced within the scope of the appended claims.

TABLE 1

	1	2	3	4	5	6	7	8	9	10
M1	G	G	C	T	C	T	G	C	G	G
M2	A	A	C	T	C	A	T	C	G	G
M3	A	A	C	T	C	A	T	C	G	G
M4	A	A	C	G	C	A	G	C	G	A
M5	G	G	G	T	C	A	G	C	G	G
M6	A	A	C	T	T	A	G	C	A	G
M7	G	G	C	T	C	A	G	C	G	G
M8	G	G	C	T	C	A	G	C	G	G
M9	A	A	C	T	T	A	G	C	A	G
M10	G	G	C	T	C	A	G	T	G	G

TABLE 2

																				SEQ ID NO:	
starting sequence	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G	1
M1	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G	1
M10	T	G	A	G	C	A	A	C	A	G	T	G	G	A	A	A	T	T	T	G	1
M2	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	2
M3	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	2
M4	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	2
M5	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	1
M6	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	2
M7	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	1
M8	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	1
M9	T	G	A	G	C	A	A	C	A	A	T	G	G	A	A	A	T	T	T	G	2

TABLE 3

																					SEQ ID NO:
starting sequence	G	C	A	G	T	T	T	G	A	G	T	G	T	C	T	C	T	G	G	T	3
M1	G	C	A	G	T	T	T	G	A	G	T	G	T	C	T	C	T	G	G	T	3
M10	G	C	A	G	T	T	T	G	A	G	T	G	T	C	T	C	T	G	G	T	3
M2	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	4
M3	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	4
M4	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	4
M5	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	3
M6	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	4
M7	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	3
M8	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	3
M9	G	C	A	G	T	T	T	G	A	A	T	G	T	C	T	C	T	G	G	T	4

TABLE 4

																					SEQ ID NO:
starting sequence	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M1	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M10	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M2	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M3	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M4	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M5	G	T	N	A	N	T	N	C	T	G	N	G	C	A	A	N	T	A	N	C	6
M6	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M7	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M8	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5
M9	G	T	A	A	A	T	G	C	T	C	T	G	C	A	A	A	T	A	A	C	5

TABLE 5

																					SEQ ID NO:	
starting sequence	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	C	C	G	G	C	7	
M1	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	N	C	C	G	N	C	8
M10	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	C	C	G	G	C	7	
M2	G	G	C	T	C	C	A	A	G	C	G	G	T	G	N	N	C	G	N	C	9	
M3	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	N	N	G	N	C	10	
M4	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	C	C	G	N	C	11	
M5	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	N	N	G	G	N	12	
M6	G	G	C	T	C	N	N	N	N	T	N	G	N	N	N	N	N	N	N	C	13	
M7	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	N	C	G	N	C	8	
M8	G	G	C	T	C	C	A	A	G	C	G	G	T	G	C	N	C	G	G	C	14	
M9	G	G	C	T	C	C	N	A	A	T	G	O	T	N	N	N	N	O	N	C	15	

TABLE 6

		SEQ ID NO:																					
starting sequence	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T			
M1	G	A	C	C	T	C	T	T	T	T	G	C	T	C	G	T	T	A	T	T		16	
M10	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		17	
M2	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M3	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M4	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M5	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M6	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M7	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M8	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	
M9	G	A	C	C	T	C	T	T	T	A	G	C	T	C	G	T	T	A	T	T		16	

15

TABLE 7

		SEQ ID NO:																					
starting sequence	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C			
M1	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M10	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M2	G	G	G	C	C	T	C	A	N	T	T	T	T	T	G	A	T	T	T	C		19	
M3	G	G	G	C	C	T	C	A	N	T	T	T	T	T	G	A	T	T	T	C		19	
M4	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M5	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M6	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M7	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M8	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	
M9	G	G	G	C	C	T	C	A	A	G	A	T	T	T	G	A	T	T	T	C		18	

TABLE 8

		SEQ ID NO:																					
starting sequence	A	G	G	G	G	G	G	C	T	T	T	T	T	C	C	A	G	C	T	C			
M1	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		20	
M10	A	G	G	G	G	G	G	C	T	T	T	T	T	C	C	A	G	C	T	C		21	
M2	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		20	
M3	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	
M4	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	
M5	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	
M6	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	
M7	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	
M8	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	
M9	A	G	G	G	G	G	G	C	T	C	T	T	T	C	C	A	G	C	T	C		21	

TABLE 9

		SEQ ID NO:																					
starting sequence	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A			
M1	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M10	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M2	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M3	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M4	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M5	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M6	G	C	C	T	C	C	T	T	N	A	T	T	C	T	A	C	G	A	C	A		23	
M7	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M8	G	C	C	T	C	C	T	T	C	G	T	T	C	T	A	C	G	A	C	A		22	
M9	G	C	C	T	C	C	T	T	C	A	T	T	C	T	A	C	G	A	C	A		24	

TABLE 10

																					SEQ ID NO:
starting sequence	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M1	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M10	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M2	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M3	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M4	A	G	N	G	T	N	C	G	N	A	T	C	C	T	C	A	C	C	T	G	26
M5	N	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	N	C	T	G	27
M6	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M7	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M8	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25
M9	A	G	G	G	T	G	C	G	C	G	T	C	C	T	C	A	C	C	T	G	25

TABLE 11

																				SEQ ID NO:	
starting sequence	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M1	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M10	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M2	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M3	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M4	A	A	C	C	A	G	A	A	T	G	T	A	T	T	T	T	G	A	G	G	29
M5	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M6	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M7	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M8	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28
M9	A	A	C	C	A	G	A	A	T	T	T	A	T	T	T	T	G	A	G	G	28

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(1 1 1) NUMBER OF SEQUENCES: 29

(2) INFORMATION FOR SEQ ID NO:1:

- (1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:1:

TGAGCAACAG TGGAAATTTG

20

(2) INFORMATION FOR SEQ ID NO:2:

- (1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:2:

TGAGCAACAA TGGAAATTTG

20

(2) INFORMATION FOR SEQ ID NO:3:

- (1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid

-continued

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GCAGTTTGAG TGTCTCTGGT

20

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GCAGTTTGAA TGTCTCTGGT

20

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:5:

GTAAATGCTC TGCAAATAAC

20

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GTNANTNCTG NGCAANTANC

20

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

(x i) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GGCTCCAAGC GGTGCCCGGC

20

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(i i) MOLECULE TYPE: DNA

-continued

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:8:
GGCTCCAAGC GGTGCNCGNC 20

(2) INFORMATION FOR SEQ ID NO:9:
(1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
(1 1) MOLECULE TYPE: DNA
(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:9:
GGCTCCAAGC GGTGNNGCNC 20

(2) INFORMATION FOR SEQ ID NO:10:
(1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
(1 1) MOLECULE TYPE: DNA
(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:10:
GGCTCCAAGC GGTGCNNGNC 20

(2) INFORMATION FOR SEQ ID NO:11:
(1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
(1 1) MOLECULE TYPE: DNA
(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:11:
GGCTCCAAGC GGTGCCCGNC 20

(2) INFORMATION FOR SEQ ID NO:12:
(1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
(1 1) MOLECULE TYPE: DNA
(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:12:
GGCTCCAAGC GGTGCNNGGN 20

(2) INFORMATION FOR SEQ ID NO:13:
(1) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 20 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
(1 1) MOLECULE TYPE: DNA
(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:13:
GGCTCNNNT NGNNNNNNNC 20

-continued

(2) INFORMATION FOR SEQ ID NO:14:

- (1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GGCTCCAAGC GGTGCNCGGC

2 0

(2) INFORMATION FOR SEQ ID NO:15:

- (1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GGCTCCNAAT GGTNNNGNC

2 0

(2) INFORMATION FOR SEQ ID NO:16:

- (1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GACCTCTTTA GCTCGTTATT

2 0

(2) INFORMATION FOR SEQ ID NO:17:

- (1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GACCTCTTTT GCTCGTTATT

2 0

(2) INFORMATION FOR SEQ ID NO:18:

- (1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:18:

GGGCCTCAAG ATTTGATTTT

2 0

(2) INFORMATION FOR SEQ ID NO:19:

- (1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid

-continued

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:19:

GGGCCTCANT ATTTGATTT C

2 0

(2) INFORMATION FOR SEQ ID NO:20:

(1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:20:

AGGGGGGCTT TTTCCAGCT C

2 0

(2) INFORMATION FOR SEQ ID NO:21:

(1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:21:

AGGGGGGCTC TTTCCAGCT C

2 0

(2) INFORMATION FOR SEQ ID NO:22:

(1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:22:

GCCTCCTTCG TTCTACGACA

2 0

(2) INFORMATION FOR SEQ ID NO:23:

(1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:23:

GCCTCCTTNA TTCTACGACA

2 0

(2) INFORMATION FOR SEQ ID NO:24:

(1) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 20 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

-continued

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:24:

GCCTCCTTCA TTCTACGACA 20

(2) INFORMATION FOR SEQ ID NO:25:

(1) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:25:

AGGGTGCGCG TCCTCACCTG 20

(2) INFORMATION FOR SEQ ID NO:26:

(1) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:26:

AGNGTNCGNA TCCTCACCTG 20

(2) INFORMATION FOR SEQ ID NO:27:

(1) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:27:

NGGGTGCGCG TCCTCANCTG 20

(2) INFORMATION FOR SEQ ID NO:28:

(1) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:28:

AACCAGAATT TATTTTGAGG 20

(2) INFORMATION FOR SEQ ID NO:29:

(1) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 20 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(1 1) MOLECULE TYPE: DNA

(* 1) SEQUENCE DESCRIPTION: SEQ ID NO:29:

AACCAGAATG TATTTTGAGG 20

What is claimed is:

1. An isolated nucleic acid segment of between 10 and 100 bases of which at least 10 contiguous bases including a polymorphic site are from a sequence selected from the group consisting of SEQ ID NOS:1-5, SEQ ID No:5 in which the C at position 10 is replaced by a G, SEQ ID No:7, SEQ ID No:7 in which the C at position 10 is replaced by a T, SEQ ID Nos:16-18, SEQ ID No:18 in which the G at position 10 is replaced by a T, SEQ ID Nos:20-22, SEQ ID Nos:24 and 25, SEQ ID No:25 in which the G at position 10 is replaced by an A, SEQ ID Nos:28 and 29, and the perfect complements thereof, wherein the polymorphic site occurs at position 10 in each of the SEQ. ID Nos.

2. The isolated nucleic acid segment of claim 1 that is DNA.

3. The isolated nucleic acid segment of claim 1 that is RNA.

4. The isolated nucleic acid segment of claim 1 that is less than 50 bases.

5. The isolated nucleic acid segment of claim 1 that is less than 20 bases.

6. The isolated nucleic acid segment of claim 1, wherein the ten contiguous bases are from a sequence selected from the group consisting of SEQ ID Nos:2 and 4, SEQ. ID. No:5 in which the C at position 10 is replaced by a G, SEQ ID No:7 in which the C at position 10 is replaced by a T, SEQ ID. No:17, SEQ ID No:18 in which the G at position 10 is replaced by a T, SEQ ID Nos:21 and 24, SEQ ID NO:25 in which the G at position 10 is replaced by an A, SEQ ID No:29, and the perfect complements thereof, wherein the polymorphic site occurs at position 10 in each of the SEQ ID NOS.

7. The isolated nucleic acid segment of claim 1, which is a probe, and wherein the polymorphic site occupies a central position of the probe.

8. The nucleic acid of claim 1, which is a primer and, wherein the polymorphic site occupies the 3' end of the primer.

9. An isolated nucleic acid fragment of a human X chromosome comprising at least 10 contiguous bases including a polymorphic site from a sequence selected from the group consisting of SEQ ID Nos:2 and 4, SEQ ID No:5 in which the C at position 10 is replaced by a G, SEQ ID No:7 in which the C at position 10 is replaced by a T, SEQ ID No:17, SEQ. ID. No:18 in which the G at position 10 is replaced by a T, SEQ ID NOS:21 and 24, SEQ ID NO:25 in which the G at position 10 is replaced by an A, SEQ ID NO:29, and the perfect complements thereof, wherein the polymorphic site occurs at position 10 in each of the SEQ ID NOS.

10. The isolated nucleic acid fragment of a human X-chromosome of claim 9, comprising a sequence selected from the group consisting of SEQ ID NOS:2 and 4, SEQ ID NO:5 in which the C at position 10 is replaced by a G, SEQ ID NO:7 in which the C at position 10 is replaced by a T, SEQ ID NO:17, SEQ. ID NO:18 in which the G at position 10 is replaced by a T, SEQ ID NOS:21 and 24, SEQ ID NO:25 in which the G at position 10 is replaced by an A, SEQ ID NO:29, and the perfect complements thereof, wherein the polymorphic site occurs at position 10 in each of the SEQ. ID NOS.

11. A method of determining a base occupying a polymorphic site in a nucleic acid, comprising:

obtaining the nucleic acid from an individual; and

determining a base occupying a polymorphic site in a sequence selected from the group consisting of SEQ ID NOS:2 and 4, SEQ ID NO:5 in which the C at position 10 is replaced by a G, SEQ ID NO:7 in which the C at position 10 is replaced by a T, SEQ ID NO:17, SEQ. ID NO:18 in which the G at position 10 is replaced by a T, SEQ ID NOS:21 and 24, SEQ ID NO:25 in which the G at position 10 is replaced by an A, SEQ ID NO:29, and the perfect complements thereof, wherein the polymorphic site occurs at position 10 in each of the SEQ ID NOS.

12. The method of claim 11, wherein the determining comprises determining a set of bases occupying a set of polymorphic sites in a set of sequences selected from the group consisting of SEQ ID NOS:2 and 4, SEQ ID NO:5 in which the C at position 10 is replaced by a G, SEQ ID NO:7 in which the C at position 10 is replaced by a T, SEQ ID NO:17, SEQ. ID NO:18 in which the G at position 10 is replaced by a T, SEQ ID NOS:21 and 24, SEQ ID NO:25 in which the G at position 10 is replaced by an A, SEQ ID NO:29, and the perfect complements thereof, wherein the polymorphic site occurs at position 10 in each of the SEQ. ID NOS.

13. The method of claim 12, wherein the nucleic acid is obtained from a plurality of individuals, and a base occupying one of the polymorphic sites is determined in each of the individuals, and the method further comprising testing each individual for the presence of a disease phenotype, and correlating the presence of the disease phenotype with the base.

* * * * *

(19)



Europäisches Patentamt

European Patent Office

Office européen des brevets



(11)

EP 0 717 113 A2

(12)

EUROPEAN PATENT APPLICATION

(43) Date of publication:
19.06.1996 Bulletin 1996/25

(51) Int. Cl.⁶: **C12Q 1/68**, G01N 33/48,
G06F 19/00

(21) Application number: 95307476.2

(22) Date of filing: 20.10.1995

(84) Designated Contracting States:
DE FR GB IT NL

(30) Priority: 21.10.1994 US 327525

(71) Applicant: AFFYMAX TECHNOLOGIES N.V.
Willemstad, Curaçao (AN)

(72) Inventors:
• Chee, Mark S.
Palo Alto, California 94306 (US)
• Wang, Chunwei
Cupertino, California 95014 (US)

• Jevons, Luis C.
Sunnyvale, California 94087 (US)
• Bernhart, Derek H.
Palo Alto, California 94301 (US)
• Lipshutz, Robert J.
Palo Alto, California 94301 (US)

(74) Representative: Nash, David Allan
Haseltine Lake & Co.
Hazlitt House
28 Southampton Buildings
Chancery Lane
London WC2A 1AT (GB)

(54) Computer-aided visualization and analysis system for nucleic acid sequence evaluation

(57) A computer system (1) for analyzing nucleic acid sequences is provided. The computer system is used to perform multiple methods for determining unknown bases by analyzing the fluorescence intensities of hybridized nucleic acid probes. The results of individual experiments may be improved by processing nucleic acid sequences together. Comparative analysis of multiple experiments is also provided by displaying reference sequences in one area (814) and sample sequences in another area (816) on a display device (3).

EP 0 717 113 A2

Description**COPYRIGHT NOTICE**

5 A portion of the disclosure of this patent document contains material which is subject to copyright protection. The copyright owner has no objection to the xerographic reproduction by anyone of the patent document or the patent disclosure in exactly the form it appears in the Patent and Trademark Office patent file or records, but otherwise reserves all copyright rights whatsoever.

10 GOVERNMENT RIGHTS NOTICE

Portions of the material in this specification arose in the course of or under contract nos. 92ER81275 (SBIR) between Affymetrix, Inc. and the Department of Energy and/or H600813-1, -2 between Affymetrix, Inc. and the National Institutes of Health.

15 BACKGROUND OF THE INVENTION

The present invention relates to the field of computer systems. More specifically, the present invention relates to computer systems for visualizing biological sequences, as well as for evaluating and comparing biological sequences.

20 Devices and computer systems for forming and using arrays of materials on a substrate are known. For example, PCT applications WO92/10588 and 95/11995, incorporated herein by reference for all purposes, describe techniques for sequencing or sequence checking nucleic acids and other materials. Arrays for performing these operations may be formed in arrays according to the methods of, for example, the pioneering techniques disclosed in U.S. Patent Nos. 5,445,934 and 5,384,261, and U.S. Patent Application No. 08/249,188, each incorporated herein by reference for all purposes.

25 According to one aspect of the techniques described therein, an array of nucleic acid probes is fabricated at known locations on a chip or substrate. A labeled nucleic acid is then brought into contact with the chip and a scanner generates an image file (also called a cell file) indicating the locations where the labeled nucleic acids bound to the chip. Based upon the image file and identities of the probes at specific locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA. Such systems have been used to form, for example, arrays of DNA that may be used to study and detect mutations relevant to cystic fibrosis, the P53 gene (relevant to certain cancers), HIV, and other genetic characteristics.

30 Improved computer systems and methods are needed to evaluate, analyze, and process the vast amount of information now used and made available by these pioneering technologies.

35 SUMMARY OF THE INVENTION

An improved computer-aided system for visualizing and determining the sequence of nucleic acids is disclosed. The computer system provides, among other things, improved methods of analyzing fluorescent image files of a chip containing hybridized nucleic acid probes in order to call bases in sample nucleic acid sequences.

40 According to one aspect of the invention, a computer system is used to identify an unknown base in a sample nucleic acid sequence by the steps of:

- inputting multiple probe intensities, each of the probe intensities being associated with a nucleic acid probe;
- 45 - the computer system comparing the multiple probe intensities where each of the probe intensities is substantially proportional to a nucleic acid probe hybridizing with at least one nucleic acid sequence; and

calling the unknown base according to the results of the comparison of the multiple probe intensities.

50 According to one specific aspect of the invention, a higher probe intensity is compared to a lower probe intensity to call the unknown base. According to another specific aspect of the invention, probe intensities of a sample sequence are compared to probe intensities of a reference sequence. According to yet another specific aspect of the invention, probe intensities of a sample sequence are compared to statistics about probe intensities of a reference sequence from multiple experiments.

55 According to another aspect of the invention, a method is disclosed of processing reference and sample nucleic acid sequences to reduce the variations between the experiments by the steps of:

- providing a plurality of nucleic acid probes;
- labeling the reference nucleic acid sequence with a first marker;
- labeling the sample nucleic acid sequence with a second marker; and

hybridizing the labeled reference and sample nucleic acid sequences at the same time.

According to another aspect of the invention, a computer system is used to identify mutations in a sample nucleic acid sequence by the steps of:

- 5 - inputting a first set of probe intensities, each of the probe intensities in said first set being associated with a nucleic acid probe and substantially proportional to the associated nucleic acid probe hybridizing with a reference nucleic acid sequence;
- inputting a second set of probe intensities, each of the probe intensities in said first set being associated with a nucleic acid probe and substantially proportional to the associated nucleic acid probe hybridizing with said sample
- 10 sequence;
- the computer system comparing probe intensities in the first set to probe intensities in the second set to select hybridization regions where the probe intensities in the first and second sets differ; and

identifying mutations according to characteristics of the selected regions.

- 15 According to yet another aspect of the invention, a computer system is used for comparative analysis and visualization of multiple sequences by the steps of:

- displaying at least one reference sequence in a first area on a display device; and
- displaying at least one sample sequence in a second area on said display device;

20

whereby a user is capable of visually comparing the multiple sequences.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the attached drawings.

25 BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system used to execute the software of the present invention;

Fig. 2 shows a system block diagram of a typical computer system used to execute the software of the present invention;

30 Fig. 3 illustrates an overall system for forming and analyzing arrays of biological materials such as DNA or RNA;

Fig. 4 is an illustration of the software for the overall system;

Fig. 5 illustrates the global layout of a chip formed in the overall system;

Fig. 6 illustrates conceptually the binding of probes on chips;

Fig. 7 illustrates probes arranged in lanes on a chip;

35 Fig. 8 illustrates a hybridization pattern of a target on a chip with a reference sequence as in Fig. 7;

Fig. 9 illustrates the high level flow of the intensity ratio method;

Fig. 10A illustrates the high level flow of one implementation of the reference method and Fig. 10B shows an analysis table for use with the reference method;

40 Fig. 11A illustrates the high level flow of another implementation of the reference method; Fig. 11B shows a data table for use with the reference method; Fig. 11C shows a graph of the normalized sample base intensities minus the normalized reference base intensities; and Fig. 11D shows other graphs of data in the data table;

Fig. 12 illustrates the high level flow of the statistical method;

Fig. 13 illustrates the pooling processing of a reference and sample nucleic acid sequence;

45 Figs. 14A and 14C show graphs of scaled fluorescent intensities of wild-type probes hybridizing with sample and reference sequences and 14B shows a hypothetical graph of fluorescent intensities of wild-type probes hybridizing with two sample sequences and a reference sequence;

Fig. 15 illustrates the high level flow of an embodiment that uses the hybridization data from than one base position to identify mutations in a sample sequence;

50 Fig. 16 illustrates the main screen and the associated pull down menus for comparative analysis and visualization of multiple experiments;

Fig. 17 illustrates an intensity graph window for a selected base;

Fig. 18 illustrates multiple intensity graph windows for selected bases;

Fig. 19 illustrates the intensity ratio method correctly calling a mutation in solutions with varying concentrations;

55 Fig. 20 illustrates the reference method correctly calling a mutant base where the intensity ratio method incorrectly called the mutant base; and

Fig. 21 illustrates the output of the ViewSeq™ program with four pretreatment samples and four posttreatment samples.

DESCRIPTION OF THE PREFERRED EMBODIMENT

CONTENTS

- 5 I. General
- II. Intensity Ratio Method
- III. Reference Method
- IV. Statistical Method
- V. Pooling Processing
- 10 VI. Comparative Analysis
- VII. Examples

I. General

15 In the description that follows, the present invention will be described in reference to a Sun Workstation in a UNIX environment. The present invention, however, is not limited to any particular hardware or operating system environment. Instead, those skilled in the art will find that the systems and methods of the present invention may be advantageously applied to a variety of systems, including IBM personal computers running MS-DOS or Microsoft Windows. Therefore, the following description of specific systems are for purposes of illustration and not limitation.

20 Fig. 1 illustrates an example of a computer system used to execute the software of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a floppy disk drive 14 and a hard drive (not shown) that may be utilized to store and retrieve software programs incorporating the present invention. Although a floppy disk 15 is shown as the removable media, other removable tangible media including CD-ROM, flash memory and
25 tape may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute the software of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 52, system memory 54, I/O controller 56, display adapter 58, serial port 62, disk 64, network
30 interface 66, and speaker 68. Disk 64 is representative of an internal hard drive, floppy drive, CD-ROM, flash memory, tape, or any other storage medium. Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 52 (i.e., a multi-processor system) or memory cache.

Arrows such as 70 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, speaker 68 could be connected to the other subsystems through a port or have an internal direct connection to central processor 52. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art.

The VLSIPS™ technology provides methods of making very large arrays of oligonucleotide probes on very small
40 chips. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated by reference for all purposes. The oligonucleotide probes on the DNA probe array are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

The present invention provides methods of analyzing hybridization intensity files for a chip containing hybridized nucleic acid probes. In a representative embodiment, the files represent fluorescence data from a biological array, but
45 the files may also represent other data such as radioactive intensity data or large molecule detection data. Therefore, the present invention is not limited to analyzing fluorescent measurements of hybridizations but may be readily utilized to analyze other measurements of hybridization.

For purposes of illustration, the present invention is described as being part of a computer system that designs a chip mask, synthesizes the probes on the chip, labels the nucleic acids, and scans the hybridized nucleic acid probes.
50 Such a system is fully described in U.S. Patent Application No. 08/249,188 which has been incorporated by reference for all purposes. However, the present invention may be used separately from the overall system for analyzing data generated by such systems.

Fig. 3 illustrates a computerized system for forming and analyzing arrays of biological materials such as RNA or DNA. A computer 100 is used to design arrays of biological polymers such as RNA or DNA. The computer 100 may be,
55 for example, an appropriately programmed Sun Workstation or personal computer or workstation, such as an IBM PC equivalent, including appropriate memory and a CPU as shown in Figs. 1 and 2. The computer system 100 obtains inputs from a user regarding characteristics of a gene of interest, and other inputs regarding the desired features of the array. Optionally, the computer system may obtain information regarding a specific genetic sequence of interest from an external or internal database 102 such as GenBank. The output of the computer system 100 is a set of chip design

computer files 104 in the form of, for example, a switch matrix, as described in PCT application WO 92/10092, and other associated computer files.

The chip design files are provided to a system 106 that designs the lithographic masks used in the fabrication of arrays of molecules such as DNA. The system or process 106 may include the hardware necessary to manufacture masks 110 and also the necessary computer hardware and software 108 necessary to lay the mask patterns out on the mask in an efficient manner. As with the other features in Fig. 3, such equipment may or may not be located at the same physical site, but is shown together for ease of illustration in Fig. 3. The system 106 generates masks 110 or other synthesis patterns such as chrome-on-glass masks for use in the fabrication of polymer arrays.

The masks 110, as well as selected information relating to the design of the chips from system 100, are used in a synthesis system 112. Synthesis system 112 includes the necessary hardware and software used to fabricate arrays of polymers on a substrate or chip 114. For example, synthesizer 112 includes a light source 116 and a chemical flow cell 118 on which the substrate or chip 114 is placed. Mask 110 is placed between the light source and the substrate/chip, and the two are translated relative to each other at appropriate times for deprotection of selected regions of the chip. Selected chemical reagents are directed through flow cell 118 for coupling to deprotected regions, as well as for washing and other operations. All operations are preferably directed by an appropriately programmed computer 119, which may or may not be the same computer as the computer(s) used in mask design and mask making.

The substrates fabricated by synthesis system 112 are optionally diced into smaller chips and exposed to marked receptors. The receptors may or may not be complementary to one or more of the molecules on the substrate. The receptors are marked with a label such as a fluorescein label (indicated by an asterisk in Fig. 3) and placed in scanning system 120. Scanning system 120 again operates under the direction of an appropriately programmed digital computer 122, which also may or may not be the same computer as the computers used in synthesis, mask making, and mask design. The scanner 120 includes a detection device 124 such as a confocal microscope or CCD (charge-coupled device) that is used to detect the locations where labeled receptor (*) has bound to the substrate. The output of scanner 120 is an image file(s) 124 indicating, in the case of fluorescein labeled receptor, the fluorescence intensity (photon counts or other related measurements, such as voltage) as a function of position on the substrate. Since higher photon counts will be observed where the labeled receptor has bound more strongly to the array of polymers, and since the monomer sequence of the polymers on the substrate is known as a function of position, it becomes possible to determine the sequence(s) of polymer(s) on the substrate that are complementary to the receptor.

The image file 124 is provided as input to an analysis system 126 that incorporates the visualization and analysis methods of the present invention. Again, the analysis system may be any one of a wide variety of computer system(s), but in a preferred embodiment the analysis system is based on a Sun Workstation or equivalent. The present invention provides various methods of analyzing the chip design files and the image files, providing appropriate output 128. The present invention may further be used to identify specific mutations in a receptor such as DNA or RNA.

Fig. 4 provides a simplified illustration of the overall software system used in the operation of one embodiment of the invention. As shown in Fig. 4, in some cases (such as sequence checking systems) the system first identifies the genetic sequence(s) or targets that would be of interest in a particular analysis at step 202. The sequences of interest may, for example, be normal or mutant portions of a gene, genes that identify heredity, or provide forensic information, or be all possible n-mers (where n represents the length of the nucleic acid). Sequence selection may be provided via manual input of text files or may be from external sources such as GenBank. At step 204 the system evaluates the gene to determine or assist the user in determining which probes would be desirable on the chip, and provides an appropriate "layout" on the chip for the probes. The chip usually includes probes that are complementary to a reference nucleic acid sequence which has a known sequence. A wild-type probe is a probe that will ideally hybridize with the reference sequence and thus a wild-type gene (also called the chip wild-type) would ideally hybridize with wild-type probes on the chip. The target sequence is substantially similar to the reference sequence except for the presence of mutations, insertions, deletions, and the like. The layout implements desired characteristics such as arrangement on the chip that permits "reading" of genetic sequence and/or minimization of edge effects, ease of synthesis, and the like.

Fig. 5 illustrates the global layout of a chip in a particular embodiment used for sequence checking applications. Chip 114 is composed of multiple units where each unit may contain different tilings for the chip wild-type sequence. Unit 1 is shown in greater detail and shows that each unit is composed of multiple cells which are areas on the chip that may contain probes. Conceptually, each unit is composed of multiple sets of related cells. As used herein, the term cell refers to a region on a substrate that contains many copies of a molecule or molecules of interest. Each unit is composed of multiple cells that may be placed in rows (or "lanes") and columns. In one embodiment, a set of five related cells includes the following: a wild-type cell 220, "mutation" cells 222, and a "blank" cell 224. Cell 220 contains a wild-type probe that is the complement of a portion of the wild-type sequence. Cells 222 contain "mutation" probes for the wild-type sequence. For example, if the wild-type probe is 3'-ACGT, the probes 3'-ACAT, 3'-ACCT, 3'-ACGT, and 3'-ACTT may be the "mutation" probes. Cell 224 is the "blank" cell because it contains no probes (also called the "blank" probe). As the blank cell contains no probes, labeled receptors should not bind to the chip in this area. Thus, the blank cell provides an area that can be used to measure the background intensity.

In one embodiment, numerous tiling processes are available including sequence tiling, block tiling, and opt-tiling, as described below. Of course a wide range of layout strategies may be used according to the invention herein, without departing from the scope of the invention. For example, the probes may be tiled on a substrate in an apparently random fashion where a computer system is utilized to keep track of the probe locations and correlate the data obtained from the substrate.

Opt-tiling is the process of tiling additional probes for suspected mutations. As a simple example of opt-tiling, suppose the wild-type target sequence is 5'-ACGTATGCA-3' and it is suspected that a mutant sequence has a possible T base mutation at the underlined base position. Suppose further that the chip will be synthesized with a "4x3" tiling strategy, meaning that probes of four monomers are used and that the monomers in position 3, counting left to right, of the probe are varied.

In opt-tiling, extra probes are tiled for each suspected mutation. The extra probes are tiled as if the mutation base is a wild-type base. The following shows the probes that may be generated for this example:

Table 1

Probe Sequences (From 3'-end) 4x3 Opt-Tiling					
Wild	TGCA	GCAT	CATA	ATAC	TACG
A sub.	TGAA	GCAT	CAAA	ATAC	TAAG
C sub.	TGCA	GCCT	CACA	ATCC	TACG
G sub.	TGGA	GCGT	CAGA	ATGC	TAGG
T sub.	TGTA	GCTT	CATA	ATTC	TATG
Wild	TGCA	GCAA	CAAA	AAAC	AACG
A sub.	TGAA	GCAA	CAAA	AAAC	AAAG
C sub.	TGCA	GCCA	CACA	AACC	AACG
G sub.	TGGA	GCGA	CAGA	AAGC	AAGG
T sub.	TGTA	GCTA	CATA	AATC	AATG

In the first "chip" above, the top row of the probes (along with one probe below each of the four wild-type probes) should bind to the target DNA sequence. However, if the target sequence has a T base mutation as suspected, the labeled mutant sequence will not bind that strongly to the probes in the columns around column 3. For example, the mutant receptor that could bind with the probes in column 2 is 5'-CGTT which may not bind that strongly to any of the probes in column 2 because there are T bases at the ends of the receptor and probes (i.e., not complementary). This often results in a relatively dark scanned area around a mutation.

Opt-tiling generates the second "chip" above to handle the suspected mutation as a wild-type base. Thus, the mutant receptor 5'-CGTT should bind strongly to the wild-type probe of column 2 (along with one probe below) and the mutation can be further detected.

Again referring to Fig. 4, at step 206 the masks for the synthesis are designed. At step 208 the software utilizes the mask design and layout information to make the DNA or other polymer chips. This software 208 will control relative translation of a substrate and the mask, the flow of desired reagents through a flow cell, the synthesis temperature of the flow cell, and other parameters. At step 210, another piece of software is used in scanning a chip thus synthesized and exposed to a labeled receptor. The software controls the scanning of the chip, and stores the data thus obtained in a file that may later be utilized to extract sequence information.

At step 212 a computer system according to the present invention utilizes the layout information and the fluorescence information to evaluate the hybridized nucleic acid probes on the chip. Among the important pieces of information obtained from probe arrays are the identification of mutant receptors and determination of genetic sequence of a particular receptor.

Fig. 6 illustrates the binding of a particular target DNA to an array of DNA probes 114. As shown in this simple example, the following probes are formed in the array (only one probe is shown for the wild-type probe):

3' -AGAACGT
 AGACCGT
 AGAGCGT
 AGATCGT

•
 •
 •

As shown, the set of probes differ by only one base so the probes are designed to determine the identity of the base at that position in the nucleic acid sequence.

When a fluorescein-labeled (or otherwise marked) target with the sequence 5'-TCTTGCA is exposed to the array, it is complementary only to the probe 3'-AGAACGT, and fluorescein will be primarily found on the surface of the chip where 3'-AGAACGT is located. Thus, for each set of probes that differ by only one base, the image file will contain four fluorescence intensities, one for each probe. Each fluorescence intensity can therefore be associated with the base of each probe that is different from the other probes. Additionally, the image file will contain a "blank" cell which can be used as the fluorescence intensity of the background. By analyzing the five fluorescence intensities associated with a specific base location, it becomes possible to extract sequence information from such arrays using the methods of the invention disclosed herein.

Fig. 7 illustrates probes arranged in lanes on a chip. A reference sequence is shown with five interrogation positions marked with number subscripts. An interrogation position is a base position in the reference sequence where the target sequence may contain a mutation or otherwise differ from the reference sequence. The chip may contain five probe cells that correspond to each interrogation position. Each probe cell contains a set of probes that have a common base at the interrogation position. For example, at the first interrogation position, I₁, the reference sequence has a base T. The wild-type probe for this interrogation position is 3'-TGAC where the base A in the probe is complementary to the base

Similarly, there are four "mutant" probe cells for the first interrogation position, I₁. The four mutant probes are 3'-TGAC, 3'-TGCC, 3'-TGGC, and 3'-TGTC. Each of the four mutant probes vary by a single base at the interrogation position. As shown, the wild-type and mutant probes are arranged in lanes on the chip. One of the mutant probes (in this case 3'-TGAC) is identical to the wild-type probe and therefore does not evidence a mutation. However, the redundancy gives a visual indication of mutations as will be seen in Fig. 8.

Still referring to Fig. 7, the chip contains wild-type and mutant probes for each of the other interrogation positions I₂-I₅. In each case, the wild-type probe is equivalent to one of the mutant probes.

Fig. 8 illustrates a hybridization pattern of a target on a chip with a reference sequence as in Fig. 7. The reference sequence is shown along the top of the chip for comparison. The chip includes a WT-lane (wild-type), an A-lane, a C-lane, a G-lane, and a T-lane (or U). Each lane is a row of cells containing probes. The cells in the WT-lane contain probes that are complementary to the reference sequence. The cells in the A-, C-, G-, and T-lanes contain probes that are complementary to the reference sequence except that the named base is at the interrogation position.

In one embodiment, the hybridization of probes in a cell is determined by the fluorescent intensity (e.g., photon counts) of the cell resulting from the binding of marked target sequences. The fluorescent intensity may vary greatly among cells. For simplicity, Fig. 8 shows a high degree of hybridization by a cell containing a darkened area. The WT-lane allows a simple visual indication that there is a mutation at interrogation position I₄ because the wild-type cell is not dark at that position. The cell in the C-lane is darkened which indicates that the mutation is from T->G (mutant probe cells are complementary so the C-cell indicates a G mutation).

In practice, the fluorescent intensities of cells near an interrogation position having a mutation are relatively dark creating "dark regions" around a mutation. The lower fluorescent intensities result because the cells at interrogation positions near a mutation do not contain probes that are perfectly complementary to the target sequence; thus, the hybridization of these probes with the target sequence is lower. For example, the relative intensity of the cells at interrogation positions I₃ and I₅ may be relatively low because none of the probes therein are complementary to the target sequence.

For ease of reference, one may call bases by assigning the bases the following codes:

Code	Group	Meaning
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T(U)	Thymine (Uracil)
M	A or C	aMino
R	A or G	puRine
W	A or T(U)	Weak interaction (2 H bonds)
Y	C or T(U)	pYrimidine
S	C or G	Strong interaction (3 H bonds)
K	G or T(U)	Keto
V	A, C or G	not T(U)
H	A, C or T(U)	not G
D	A, G or T(U)	not C
B	C, G or T(U)	not A
N	A, C, G, or T(U)	Insufficient intensity to call
X	A, C, G, or T(U)	Insufficient discrimination to call

Most of the codes conform to the IUPAC standard. However, code N has been redefined and code X has been added.

II. Intensity Ratio Method

The intensity ratio method is a method of calling bases in a sample nucleic acid sequence. The intensity ratio method is most accurate when there is good discrimination between the fluorescence intensities of hybrid matches and hybrid mismatches. If there is insufficient discrimination, the intensity ratio method assigns a corresponding ambiguity code to the unknown base.

For simplicity, the intensity ratio method will be described as being used to identify one unknown base in a sample nucleic acid sequence. In practice, the method is used to identify many or all the bases in a nucleic acid sequence.

The unknown base will be identified by evaluation of up to four mutation probes and a "blank" cell, which is a location where a labeled receptor should not bind to the chip since no probe is present. For example, suppose a DNA sequence of interest or target sequence contains the sequence 5'-AGAACCTGC-3' with a possible mutation at the underlined base position. Suppose that 5-mer probes are to be synthesized for the target sequence. A representative wild-type probe of 5'-TTGGA is complementary to the region of the sequence around the possible mutation. The "mutation" probes will be the same as the wild-type probe except for a different base at the third position as follows: 3'-TTAGA, 3'-TTCGA, 3'-TTGGA, and 3'-TTTGA.

If the fluorescently marked sample sequence is exposed to the above four mutation probes, the intensity should be highest for the probe that binds most strongly to the sample sequence. Therefore, if the probe 3'-TTTGA shows the highest intensity, the unknown base in the sample will generally be called an A mutation because the probes are complementary to the sample sequence.

The mutation probes are identical to the wild-type probes except that they each contain one of the four A, C, G, or T "mutations" for the unknown base. Although one of the "mutation" probes will be identical to the wild-type probe, such redundant probes are intentionally synthesized for quality control and design consistency.

The identity of the unknown base is preferably determined by evaluating the relative fluorescence intensities of up to four of the mutation probes, and the "blank" cell. Because each mutation probe is identifiable by the mutation base, a mutation probe's intensity will be referred to as the "base intensity" of the mutation base.

As a simple example of the intensity ratio method, suppose a gene of interest (target) is an HIV protease gene with the sequence 5'-ATGTGGACAGTTGTA-3' (SEQ ID NO:1). Suppose further that a sample sequence is suspected to

have the same sequence as the target sequence except for a mutation of base C to base T at the underlined base position. Although hundreds of probes may be synthesized on the chip, the complementary mutation probes synthesized to detect a mutation in the sample sequence at the suspected mutation position may be as follows:

3'-TATC
3'-TCTC
3'-TGTC (wild-type)
3'-TTTC

The mutation probe 3'-TGTC is also the wild-type probe as it should bind most strongly with the target sequence.

After the sample sequence is labeled, hybridized on the chip, and scanned, suppose the following fluorescence intensities were obtained:

3'-TATC -> 45
3'-TCTC -> 8
3'-TGTC -> 32
3'-TTTC -> 12

where the intensity is measured by the photon count detected by the scanner. The "blank" cell had a fluorescence intensity of 2. The photon counts in the examples herein are representative (not actual data) and provided for illustration purposes. In practice, the actual photon counts will vary greatly depending on the experiment parameters and the scanner utilized.

Although each fluorescence intensity is from a probe, the probes may be characterized by their unique mutation base so the bases may be said to have the following intensities:

A -> 45
C -> 8
G -> 32
T -> 12

Thus, base A will be described as having an intensity of 45, which corresponds to the intensity of the mutation probe with the mutation base A.

Initially, each mutation base intensity is reduced by the background or "blank" cell intensity. This is done as follows:

$$A \rightarrow 45 - 2 = 43$$

$$C \rightarrow 8 - 2 = 6$$

$$G \rightarrow 32 - 2 = 30$$

$$T \rightarrow 12 - 2 = 10$$

Then, the base intensities are sorted in descending order of intensity. The above bases would be sorted as follows:

A -> 43
G -> 30
T -> 10
C -> 6

Next, the highest intensity base is compared to the second highest intensity base. Thus, the ratio of the intensity of base A to the intensity of base G is calculated as follows: $A:G = 43 / 30 = 1.4$. The ratio A:G is then compared to a predetermined ratio cutoff, which is a number that specifies the ratio required to identify the unknown base. For example, if the ratio cutoff is 1.2, the ratio A:G is greater than the ratio cutoff ($1.4 > 1.2$) and the unknown base is called by the mutation probe containing the mutation A. As probes are complementary to the sample sequence, the sample sequence is called as having a mutation T, resulting in a called sample sequence of 5'-ATGTGGATAGTTGTA-3' (SEQ ID NO:2).

As another example, suppose everything else is the same as in the previous example except that the sorted background adjusted intensities were as follows:

C -> 42
A -> 40
G -> 10
T -> 8

The ratio of the highest intensity base to the second highest intensity base (C:A) is 1.05. Because this ratio is not greater than the ratio cutoff of 1.2, the unknown base will be called as being ambiguously one of two or more bases as follows.

The second highest intensity base is then compared to the third highest base. The ratio of A:G is 4. The ratio of A:G is then compared to the ratio cutoff of 1.2. As the ratio A:G is greater than the ratio cutoff ($4 > 1.2$), the unknown base is called by the mutation probes containing the mutations C or A. As probes are complementary to the sample sequence,

the sample sequence is called as having either a mutation G or T, resulting in a sample sequence of 5'-ATGTGGAK-AGTTGTA-3' (SEQ ID NO:3) where K is the IUPAC code for G or T(U).

The ratio cutoff in the previous examples was equal to 1.2. However, the ratio cutoff will generally need to be adjusted to produce optimal results for the specific chip design and wild-type target. Also, although the ratio cutoff used has been the same for each ratio comparison, the ratio cutoff may vary depending on whether the ratio comparisons involve the highest, second highest, third highest, etc. intensity base.

Fig. 9 illustrates the high level flow of the intensity ratio method. At step 302 the four base intensities are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers in future calculations.

At step 304 the base intensities are sorted by intensity. Each base is then associated with a number from 1 to 4. The base with the highest intensity is 1, second highest 2, third highest 3, and fourth highest 4. Thus, the intensity of base 1 \geq base 2 \geq base 3 \geq base 4.

At step 306 the highest intensity base (base 1) is checked to see if it has sufficient intensity to call the unknown base. The intensity is checked by determining if the intensity of base 1 is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the intensity a base intensity must be over the background intensity in order to correctly call the unknown base. Thus, the background adjusted base intensity must be greater than the background difference cutoff or the unknown is not callable.

If the intensity of base 1 is not greater than the background difference cutoff, the unknown base is assigned the code N (insufficient intensity) as shown at step 308. Otherwise, the ratio of the intensity of base 1 to base 2 is calculated as shown at step 310.

At step 312 the ratio of intensity of bases 1:2 is compared to the ratio cutoff. If the ratio 1:2 is greater than the ratio cutoff, the unknown base is called as the complement of the highest intensity base (base 1) as shown at step 314. Otherwise, the ratio of the intensity of base 2 to base 3 is calculated as shown at step 316.

At step 318 the ratio of intensity of bases 2:3 is compared to the ratio cutoff. If the ratio 2:3 is greater than the ratio cutoff, the unknown base is called as being an ambiguity code specifying the complements of the highest or second highest intensity bases (base 1 or 2) as shown at step 320. Otherwise, the ratio of the intensity of base 3 to base 4 is calculated as shown at step 322.

At step 324 the ratio of intensity of bases 3:4 is compared to the ratio cutoff. If the ratio 3:4 is greater than the ratio cutoff, the unknown base is called as being an ambiguity code specifying the complements of the highest, second highest, or third highest bases (base 1, 2 or 3) as shown at step 326. Otherwise, the unknown base is assigned the code X (insufficient discrimination) as shown at step 328.

The advantage of the intensity ratio method is that it is very accurate when there is good discrimination between the fluorescence intensities of hybrid matches and hybrid mismatches. However, if the base corresponding to a correct hybrid gives a lower intensity than a mismatch (e.g., as a result of cross-hybridization), incorrect identification of the base will result. For this reason, however, the method is useful for comparative assessment of hybridization quality and as an indicator of sequence-specific problem spots. For example, the intensity ratio method has been used to determine that ambiguities and miscalls tend to be very different from sequence to sequence, and reflect predominantly the composition and repetitiveness of the sequence. It has also been used to assess improvements obtained by varying hybridization conditions, sample preparation, and post-hybridization treatments (e.g., RNase treatment).

III. Reference Method

The reference method is a method of calling bases in a sample nucleic acid sequence. The reference method depends very little on discrimination between the fluorescence intensities of hybrid matches and hybrid mismatches, and therefore is much less sensitive to cross-hybridization. The method compares the probe intensities of a reference sequence to the probe intensities of a sample sequence. Any significant changes are flagged as possible mutations. There are two implementations of the reference method disclosed herein.

For simplicity, the reference method will be described as being used to identify one unknown base in a sample nucleic acid sequence. In practice, the method is used to identify many or all the bases in a nucleic acid sequence.

The unknown base will be called by comparing the probe intensities of a reference sequence to the probe intensities of a sample sequence. Preferably, the probe intensities of the reference sequence and the sample sequence are from chips having the same chip wild-type. However, the reference sequence may or may not be exactly the same as the chip wild-type, as it may have mutations.

The bases at the same position in the reference and sample sequences will each be associated with up to four mutation probes and a "blank" cell. The unknown base in the sample sequence is called by comparing probe intensities of the sample sequence to probe intensities of the reference sequence. For example, suppose the chip wild-type contains the sequence 5'-AGACCTTGC-3' and it is suspected that the sample has a possible mutation at the underlined base

position, which is the unknown base that will be called by the reference method. The "mutation" probes for the sample sequence may be as follows: 3'-GAAA, 3'-GCAA, 3'-GGAA, and 3'-GTAA, where 3'-GGAA is the wild-type probe.

Suppose further that a reference sequence, which differs from the chip wild-type by one base mutation, has the sequence 5'-AGACATTGC-3' where the mutation base is underlined. The "mutation" probes for the reference sequence may be as follows: 3'-TGAAA, 3'-TGCAA, 3'-TGGA, and 3'-TGTA, where 3'-TGTA is the reference wild-type probe since the reference sequence is known. Although generally the sample and reference sequences were tiled with the same chip wild-type, this is not required, and the tiling methods do not have to be identical as shown by the use of two probe lengths in the example. Thus, the unknown base will be called by comparing the "mutation" probes of the sample sequence to the "mutation" probes of the reference sequence. As before, because each mutation probe is identifiable by the mutation base, the mutation probes' intensities will be referred to as the "base intensities" of their respective mutation bases.

As a simple example of one implementation of the reference method, suppose a gene of interest (target) has the sequence 5'-AA~~AA~~CTGAAAA-3' (SEQ ID NO:4). Suppose a reference sequence has the sequence 5'-AAAACCGAAAA-3' (SEQ ID NO:5), which differs from the target sequence by the underlined base. The reference sequence is marked and exposed to probes on a chip with the target sequence being the chip wild-type. Suppose further that a sample sequence is suspected to have the same sequence as the target sequence except for a mutation at the underlined base position in 5'-AAAACIGAAAA-3' (SEQ ID NO:4). The sample sequence is also marked and exposed to probes on a chip with the target sequence being the chip wild-type. After hybridization and scanning, the following probe intensities (not actual data) were found for the respective complementary probes:

Reference	Sample
3'-TGAC -> 12	3'-GACT -> 11
3'-TGCC -> 9	3'-GCCT -> 30
3'-TGGC -> 80	3'-GGCT -> 60
3'-TGTC -> 15	3'-GTCT -> 6

Although each fluorescence intensity is from a probe, the probes may be identified by their unique mutation base so the bases may be said to have the following intensities:

Reference	Sample
A -> 12	A -> 11
C -> 9	C -> 30
G -> 80	G -> 60
T -> 15	T -> 6

Thus, base A of the reference sequence will be described as having an intensity of 12, which corresponds to the intensity of the mutation probe with the mutation base A. The reference method will now be described as calling the unknown base in the sample sequence by using these intensities.

Fig. 10A illustrates the high level flow of one implementation of the reference method. For illustration purposes, the reference method is described as filling in the columns (identified by the numbers along the bottom) of the analysis table shown in Fig. 10B. However, the generation of an analysis table is not necessary to practice the method. The analysis table is shown to aid the reader in understanding the method.

At step 402 the four base intensities of the reference and sample sequences are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Each set of "mutation" probes has an associated "blank" cell. Suppose that the reference "blank" cell intensity is 1 and the sample "blank" cell intensity is 2. The base intensities are then background subtracted as follows:

Reference	Sample
A -> $12 - 1 = 11$	A -> $11 - 2 = 9$
C -> $9 - 1 = 8$	C -> $30 - 2 = 28$
G -> $80 - 1 = 79$	G -> $60 - 2 = 58$
T -> $15 - 1 = 14$	T -> $6 - 2 = 4$

Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers in future calculations.

For identification, the position of each base of interest in the reference and sample sequences is placed in column 1 of the analysis table. Also, since the reference sequence is a known sequence, the base at this position is known and is referred to as the reference wild-type. The reference wild-type is placed in column 2 of the analysis table, which is C for this example.

At step 404 the base intensity associated with the reference wild-type (column 2 of the analysis table) is checked to see if it has sufficient intensity to call the unknown base. In this example, the reference wild-type is C. However, the base intensity associated with the wild-type is the G base intensity, which is 79 in this example. This is because the base intensities actually represent the complementary "mutation" probes. The G base intensity is checked by determining if its intensity is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the intensity the base intensities must be above the background intensity in order to correctly call the unknown base. Thus, the base intensity associated with the reference wild-type must be greater than the background difference cutoff or the unknown base is not callable.

If the background difference cutoff is 5, the base intensity associated with the reference wild-type has sufficient intensity ($79 > 5$) so a P (pass) is placed in column 3 of the analysis table as shown at step 406. Otherwise, at step 407 an F (fail) is placed in column 3 of the analysis table.

At step 408 the ratio of the base intensity associated with the reference wild-type to each of the possible bases are calculated. The ratio of the base intensity associated with the reference wild-type to itself will be 1 and the other ratios will usually be greater than 1. The base intensity associated with the reference wild-type is G so the following ratios are calculated:

$$G:A \rightarrow 79 / 11 = 7.2$$

$$G:C \rightarrow 79 / 8 = 9.9$$

$$G:G \rightarrow 79 / 79 = 1.0$$

$$G:T \rightarrow 79 / 14 = 5.6$$

These ratios are placed in columns 4 through 7 of the analysis table, respectively.

At step 410 the highest base intensity associated with the sample sequence is checked to see if it has sufficient intensity to call the unknown base. The highest base intensity is checked by determining if the intensity is greater than the background difference cutoff. Thus, the highest base intensity must be greater than the background difference cutoff or the unknown base is not callable.

Again, if the background difference cutoff is 5, the highest base intensity, which is G in this example, has sufficient intensity ($58 > 5$) so a P (pass) is placed in column 8 of the analysis table as shown at step 412. Otherwise, at step 413 an F (fail) is placed in column 8 of the analysis table.

At step 414 the ratios of the highest base intensity of the sample to each of the possible bases are calculated. The ratio of the highest base intensity to itself will be 1 and the other ratios will usually be greater than 1. Thus, the highest base intensity is G so the following ratios are calculated:

$$G:A \rightarrow 58 / 9 = 6.4$$

$$G:C \rightarrow 58 / 28 = 2.3$$

$$G:G \rightarrow 58 / 58 = 1.0$$

$$G:T \rightarrow 58 / 4 = 14.5$$

These ratios are placed in columns 9 through 12 of the analysis table, respectively.

At step 416 if both the reference and sample sequence probes failed to have sufficient intensity to call the unknown base, meaning there is an 'F' in columns 3 and 8 of the analysis table, the unknown base is assigned the code N (insufficient intensity) as shown at step 418. An 'N' is placed in column 17 of the analysis table. Additionally, a confidence code of 9 is placed in column 18 of the analysis table where the confidence codes have the following meanings:

Code	Meaning
0	Probable reference wild-type
1	Probable mutation
2	Reference sufficient intensity, insufficient intensity in sample suggests possible mutation
3	Borderline differences, unknown base ambiguous
4	Sample sufficient intensity, insufficient intensity in reference to allow comparison
5-8	Currently unassigned
9	Insufficient intensity in reference and sample, no interpretation possible

The confidence codes are useful for indicating to the user the resulting analysis of the reference method.

At step 420 if only the reference sequence probes failed to have sufficient intensity to call the unknown base, meaning there is an 'F' in column 3 and a 'P' in column 8 of the analysis table, the unknown base is assigned the code N (insufficient intensity) as shown at step 422. An 'N' is placed in column 17 and a confidence code of 4 is placed in column 18 of the analysis table.

At step 424 if only the sample sequence probes failed to have sufficient intensity to call the unknown base, meaning there is a 'P' in column 3 and a 'F' in column 8 of the analysis table, the unknown base is assigned the code N (insufficient intensity) as shown at step 426. An 'N' is placed in column 17 and a confidence code of 2 is placed in column 18 of the analysis table.

In this example, both the reference and sample sequence probes have sufficient intensity to call the unknown base. At step 428 the ratios of the reference ratios to the sample ratios for each base type are calculated. Thus, the ratio A:A (column 4 to column 9) is placed in column 13 of the analysis table. The ratio C:C (column 5 to column 10) is placed in column 14 of the analysis table. The ratio G:G (column 6 to column 11) is placed in column 15 of the analysis table. Lastly, the ratio T:T (column 7 to column 12) is placed in column 16 of the analysis table. These ratios are calculated as follows:

$$A:A \rightarrow 7.2 / 6.4 = 1.1$$

$$C:C \rightarrow 9.9 / 2.3 = 4.3$$

$$G:G \rightarrow 1.0 / 1.0 = 1.0$$

$$T:T \rightarrow 5.6 / 14.5 = 0.4$$

The unknown base is called by comparing these ratios of ratios to two predetermined values as follows.

At step 430 if all the ratios of ratios (columns 13 to 16 of the analysis table) are less than a predetermined lower ratio cutoff, the unknown base is assigned the code of the reference wild-type as shown at step 432. Thus, the code for the reference wild-type (as shown in column 2) would be placed in column 17 and a confidence code of 0 would be placed in column 18 of the analysis table.

At step 434 if all the ratios of ratios are less than a predetermined upper ratio cutoff, the unknown base is assigned an ambiguity code that indicates the unknown base may be any one of the bases that has a complementary ratio of ratios greater than the lower ratio cutoff and less than the upper ratio cutoff as shown at step 436. Thus, if the ratio of ratios for A:A, C:C and G:G are all greater than the lower ratio cutoff and less than the upper ratio cutoff, the unknown base would be assigned the code B (meaning "not A"). This is because the ratios of ratios are complementary to their respective base as follows:

A:A -> T

C:C -> G

G:G -> C

so the unknown base would be called as being either C, G, or T, which is identified by the IUPAC code B. This ambiguity code would be placed in column 17 and a confidence code of 3 would be placed in column 18 of the analysis table.

At step 438 at least one of the ratios of ratios is greater than the upper ratio cutoff and the unknown base is called as the base complementary to the highest ratio of ratios. The code for the base complementary to the highest ratio of ratios would be placed in column 17 and a confidence code of 1 would be placed in column 18 of the analysis table.

Assume for the purposes of this example that the lower ratio cutoff is 1.5 and the upper ratio cutoff is 3. Again, the ratios of ratios are as follows:

A:A -> 1.1

C:C -> 4.3

G:G -> 1.0

T:T -> 0.4

As all the ratios of ratios are not less than the upper ratio cutoff, the unknown base is called the base complementary to the highest ratio of ratios. The highest ratio of ratios is C:C, which has a complementary base G. Thus, the unknown base is called G which is placed in column 17 and a confidence code of 1 is placed in column 18 of the analysis table.

The example shows how the unknown base in the sample nucleic acid sequence was correctly called as base G. Although the complementary "mutation" probe associated with the base G (3'-GCCT) did not have the highest fluorescence intensity, the unknown base was called as base G because the associated "mutation" probe had the highest ratio increase over the other "mutation" probes.

Fig. 11A illustrates the high level flow of another implementation of the reference method. As in the previous implementation, this implementation also compares the probe intensities of a reference sequence to the probe intensities of a sample sequence. However, this implementation differs conceptually from the previous implementation in that neighboring probe intensities are also analyzed, resulting in more accurate base calling.

As a simple example of this implementation of the reference method, suppose a reference sequence has a sequence of 5'-AAACCCAATCCACATCA-3' (SEQ ID NO:6) and a sample sequence has a sequence of 5'-AAACCCAGTCCACATCA-3' (SEQ ID NO:7), where the mutant base is underlined. Thus, there is a mutation of A to G. Suppose further that the reference and sample sequences are tiled on chips with the reference sequence being the chip wild-type. This implementation of the reference method will be described as identifying this mutation base.

For illustration purposes, this implementation of the reference method is described as filling in a data table shown in Fig. 11B (SEQ ID NO:6, SEQ ID NO:28, SEQ ID NO:29). Although the data table contains more data than is required for this implementation, the portions of the data table that are produced by steps in Fig. 11A are shown with the same reference numerals. The generation of a data table is not necessary, however, and is shown to aid the reader in understanding the method. The mutant base position is at position 241 in the reference and sample sequences, which is shown in bold in the data table.

At step 502 the base intensities of the reference and sample sequences are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers. In the data table, data 502A is the background subtracted base intensities for the reference sequence and data 502B is the background subtracted base intensities for the sample sequence (also called the "mutant" sequence in the data table).

At step 504 the base intensity associated with the reference wild-type is checked to see if it has sufficient intensity to call the unknown base. In this example, the reference wild-type is base A at position 241. The base intensity associated with the reference wild-type is identified by a lower case "a" in the left hand column. Thus, the base intensities in the data table are not identified by their complements and the reference wild-type at the mutation position has an intensity of 385. The reference wild-type intensity of 385 is checked by determining if its intensity is greater than a predetermined background difference cutoff. The background difference cutoff is a number that specifies the intensity the base intensities must be over the background intensity in order to correctly call the unknown base. Thus, the base intensity associated with the reference wild-type must be greater than the background difference cutoff or the unknown base is not callable.

If the base intensity associated with the reference wild-type is not greater than the background difference cutoff, the wild-type sequence would fail to have sufficient intensity as shown at step 506. Otherwise, at step 508 the wild-type sequence would pass by having sufficient intensity.

At step 510 calculations are performed on the background subtracted base intensities of the reference sequence in order to "normalize" the intensities. Each position in the reference sequence has four background subtracted base intensities associated with it. The ratio of the intensity of each base to the sum of the intensities of the possible bases (all four) is calculated, resulting in four ratios, one for each base as shown in the data table. Thus, the following ratios would be calculated at each position in the reference sequence:

$$A \text{ ratio} = A / (A + C + G + T)$$

$$C \text{ ratio} = C / (A + C + G + T)$$

$$G \text{ ratio} = G / (A + C + G + T)$$

$$T \text{ ratio} = T / (A + C + G + T)$$

At position 241, A ratio would be the wild-type ratio. These ratios are generally calculated in order to "normalize" the intensity data as the photon counts may vary widely from experiment to experiment. Thus, the ratios provide a way of reconciling the intensity variations across experiments. Preferably, if the photon counts do not vary widely from experiment to experiment, the probe intensities do not need to be "normalized."

At step 512 the highest base intensity associated with the sample sequence is checked to see if it has sufficient intensity to call the unknown base. The intensity is checked by determining if the highest intensity sample base is greater than the background difference cutoff. If the intensity is not greater than the background difference cutoff, the sample sequence fails to have sufficient intensity as shown at step 514. Otherwise, at step 516 the sample sequence passes by having sufficient intensity.

At step 518 calculations are performed on the background subtracted base intensities of the sample sequence in order to "normalize" the intensities. Each position in the sample sequence has four background subtracted base intensities associated with it. The ratios of the intensity of each base to the sum of the intensities of the possible bases (all four) are calculated, resulting in four ratios, one for each base as shown in the data table.

At step 520 if either the reference or sample sequences failed to have sufficient intensity, the unknown base is assigned the code N (insufficient intensity) as shown at step 522.

At step 524 the normalized base intensities of the reference sequence are subtracted from the normalized base intensities of the sample sequence. Thus, at each position the following calculations are performed:

$$A \text{ Difference} = \text{Sample A Ratio} - \text{Reference A Ratio}$$

$$C \text{ Difference} = \text{Sample C Ratio} - \text{Reference C Ratio}$$

$$G \text{ Difference} = \text{Sample G Ratio} - \text{Reference G Ratio}$$

$$T \text{ Difference} = \text{Sample T Ratio} - \text{Reference T Ratio}$$

where the reference and sample ratios are calculated at steps 510 and 518, respectively. The base differences resulting from these calculations are shown in the data table.

At step 526 each position is checked to see if there is a base difference greater than an upper difference cutoff and a base difference lower than a lower difference cutoff. For example, Fig. 11C shows a graph the normalized sample base intensities minus the normalized reference base intensities. Suppose that the upper difference cutoff is 0.15 and the lower difference cutoff is -0.15 as shown by the horizontal lines in Fig. 11C. At the mutation position (labeled with a reference 0), the G difference is 0.28 which is greater than 0.15, the upper difference cutoff. Similarly, the A difference is -0.32 which is less than -0.15, the lower difference cutoff. As there is a base difference above the upper difference cutoff and a base difference below the lower difference cutoff, there may be mutation at this position.

If there is neither a base difference above the upper difference cutoff nor a base difference below the lower difference cutoff, the base at that position is assigned the code of the reference wild-type base as shown at step 528.

At step 530 the ratio of the highest background subtracted base intensity in the sample to the background subtracted reference wild-type base intensity is calculated. For example, at the mutation position 241 in the data table, the highest background subtracted base intensity in the sample is 571 (base G). The background subtracted reference wild-type base intensity is 385 (base A). The ratio of 571:385 is calculated and results in 1.48 as shown in the data table.

At step 532 these ratios are compared to a ratio at a neighboring position. The ratio for the n^{th} position is subtracted from the ratio for the r^{th} position, where $r = n + 1$. For example, at the mutation position 241 in the data table, the ratio at position 242 (which equals 1.02) is subtracted from the ratio at position 241 (which equals 1.48). It has been found that a mutant can be confidently detected by analyzing the difference of these neighboring ratios.

Fig. 11D shows other graphs of data in the data table. Of particular importance is the graph identified as 532 because this is a graph of the calculations at step 532. The pattern shown in a box in graph 532 has been found to be characteristic of a mutation. Thus, if this pattern is detected, the base is called as the base (or bases) with a normalized difference greater than the upper difference cutoff as shown at step 536. For example, the pattern was detected and at step 526 it was shown that base G had a normalized difference of 0.28, which is greater than the upper difference cutoff of 0.15. Therefore, the base at position 241 in the sample sequence is called a base G, which is a mutation from the reference sequence (A to G).

If the pattern is not detected at step 534, the base at that position is assigned the code of the reference wild-type base as shown at step 538.

This second implementation of the reference method is preferable in some instances as it takes into account probe intensities of neighboring probes. Thus, the first implementation may not have detected the A to G mutation in this example.

The advantage of the reference method is that the correct base can be called even in the presence of significant levels of cross-hybridization, as long as ratios of intensities are fairly consistent from experiment to experiment. In practice, the number of miscalls and ambiguities is significantly reduced, while the number of correct calls is actually increased, making the reference method very useful for identifying candidate mutations. The reference method has also been used to compare the reproducibility of experiments in terms of base calling.

IV. Statistical Method

The statistical method is a method of calling bases in a sample nucleic acid sequence. The statistical method utilizes the statistical variation across experiments to call the bases. Therefore, the statistical method is preferable when data from multiple experiments is available and the data is fairly consistent across the experiments. The method compares the probe intensities of a sample sequence to statistics of probe intensities of a reference sequence in multiple experiments.

For simplicity, the statistical method will be described as being used to identify one unknown base in a sample nucleic acid sequence. In practice, the method is used to identify many or all the bases in a nucleic acid sequence.

The unknown base will be called by comparing the probe intensities of a sample sequence to statistics on probe intensities of a reference sequence in multiple experiments. Generally, the probe intensities of the sample sequence and the reference sequence experiments are from chips having the same chip wild-type. However, the reference sequence may or may not be equal to the chip wild-type, as it may have mutations.

A base at the same position in the reference and sample sequences will be associated with up to four mutation probes and a "blank" cell. As before, because each mutation probe is identifiable by the mutation base, the mutation probes' intensities will be referred to as the "base intensities" of their respective mutation bases.

As a simple example of the statistical method, suppose a gene of interest (target) has the sequence 5'-AAAAC-TGAAAA-3' (SEQ ID NO:4). Suppose a reference sequence has the sequence 5'-AAAACCGAAAA-3' (SEQ ID NO:5), which differs from the target sequence by the underlined base. Suppose further that a sample sequence is suspected to have the same sequence as the target sequence except for a T base mutation at the underlined base position in 5'-AAAACTGAAAA-3' (SEQ ID NO:4). Suppose that in multiple experiments the reference sequence is marked and exposed to probes on a chip. Suppose further the sample sequence is also marked and exposed to probes on a chip.

The following are complementary "mutation" probes that could be used for a reference experiment and the sample sequence:

Reference	Sample
3'-TGAC	3'-GACT
3'-TGCC	3'-GCCT
3'-TGGC	3'-GGCT
3'-TGTC	3'-GTCT

The "mutation" probes shown for the reference sequence may be from only one experiment, the other experiments may

have different "mutation" probes, chip wild-types, tiling methods, and the like. Although each fluorescence intensity is from a probe, since the probes may be identified by their unique mutation bases, the probe intensities may be identified by their respective bases as follows:

Reference	Sample
3'-TGAC -> A	3'-GACT -> A
3'-TGCC -> C	3'-GCCT -> C
3'-TGGC -> G	3'-GGCT -> G
3'-TGTC -> T	3'-GTCT -> T

Thus, base A of the reference sequence will be described as having an intensity which corresponds to the intensity of the mutation probe with the mutation base A. The statistical method will now be described as calling the unknown base in the sample sequence by using this example.

Fig. 12 illustrates the high level flow of the statistical method. At step 602 the four base intensities associated with the sample sequence and each of the multiple reference experiments are adjusted by subtracting the background or "blank" cell intensity from each base intensity. Preferably, if a base intensity is then less than or equal to zero, the base intensity is set equal to a small positive number to prevent division by zero or negative numbers.

At step 604 the intensities of the reference wild-type bases in the multiple experiments are checked to see if they all have sufficient intensity to call the unknown base. The intensities are checked by determining if the intensity of the reference wild-type base of an experiment is greater than a predetermined background difference cutoff. The wild-type probe shown earlier for the reference sequence is 3'-TGGC, and thus the G base intensity is the wild-type base intensity. These steps are analogous to steps in the other two methods described herein.

If the intensity of any one of the reference wild-type bases is not greater than the background difference cutoff, the wild-type experiments fail to have sufficient intensity as shown at step 606. Otherwise, at step 608 the wild-type experiments pass by having sufficient intensity.

At step 610 calculations are performed on the background subtracted base intensities of each of the reference experiments in order to "normalize" the intensities. Each reference experiment has four background subtracted base intensities associated with it: one wild-type and three for the other possible bases. In this example, the G base intensity is the wild-type, the A, C, and T base intensities being the "other" intensities. The ratios of the intensity of each base to the sum of the intensities of the possible bases (all four) are calculated, giving one wild-type ratio and three "other" ratios. Thus, the following ratios would be calculated:

$$A \text{ ratio} = A / (A + C + G + T)$$

$$C \text{ ratio} = C / (A + C + G + T)$$

$$G \text{ ratio} = G / (A + C + G + T)$$

$$T \text{ ratio} = T / (A + C + G + T)$$

where G ratio is the wild-type ratio and A, C, and T ratios are the "other" ratios. These four ratios are calculated for each reference experiment. Thus if the number of reference experiments is n, there would be 4n ratios calculated. These ratios are generally calculated in order to "normalize" the intensity data, as the photon counts may vary widely from experiment to experiment. However, if the probe intensities do not vary widely from experiment to experiment, the probe intensities do not need to be "normalized."

At step 612 statistics are prepared for the ratios calculated for each of the reference experiments. As stated before, each reference experiment will be associated with one wild-type ratio and three "other" ratios. The mean and standard deviation are calculated for all the wild-type ratios. The mean and standard deviation are also calculated for each of the other ratios, resulting in three other means and standard deviations for each of the bases that is not the wild-type base. Therefore, the following would be calculated:

Mean and standard deviation of A ratios

Mean and standard deviation of C ratios

Mean and standard deviation of G ratios

Mean and standard deviation of T ratios

5 where the mean and standard deviation of the G ratios are also known as the wild-type mean and the wild-type standard deviation, respectively. The mean and standard deviation of the A, C, and T means and standard deviations are also known collectively as the "other" means and standard deviations.

Suppose that the preceding calculations produced the following data:

10 A ratios -> mean = 0.16 std. dev. = 0.003

C ratios -> mean = 0.03 std. dev. = 0.002

G ratios -> mean = 0.71 std. dev. = 0.050

15 T ratios -> mean = 0.11 std. dev. = 0.004

20 In one embodiment, the steps up to and including step 612 are performed in a preprocessing stage for the multiple wild-type experiments. The results of the preprocessing stage are stored in a file so that the reference calculations do not have to be repeatedly calculated, improving performance.

At step 614 the highest base intensity associated with the sample sequence is checked to see if it has sufficient intensity to call the unknown base. The intensity is checked by determining if the highest intensity unknown base is greater than the background difference cutoff. If the intensity is not greater than the background difference cutoff, the sample sequence fails to have sufficient intensity as shown at step 616. Otherwise, at step 618 the sample sequence passes by having sufficient intensity.

25 At step 620 calculations are performed on the four background subtracted intensities of the sample sequence. The ratios of the background subtracted intensity of each base to the sum of the background subtracted intensities of the possible bases (all four) are calculated, giving four ratios, one for each base. For consistency, the ratio associated with the reference wild-type base is called the wild-type ratio, with there being three "other" ratios. Thus, the following ratios are calculated:

$$A \text{ ratio} = A / (A + C + G + T)$$

$$C \text{ ratio} = C / (A + C + G + T)$$

35 $G \text{ ratio} = G / (A + C + G + T)$

$$T \text{ ratio} = T / (A + C + G + T)$$

40 where ratio G is the wild-type ratio and ratios A, C, and T are the "other" ratios.

Suppose the background subtracted intensities associated with the sample are as follows:

A -> 310

C -> 50

G -> 26

45 T -> 100

Then, the corresponding ratios would be as follows:

$$A \text{ ratio} = 310 / (310 + 50 + 26 + 100) = 0.64$$

50 $C \text{ ratio} = 50 / (310 + 50 + 26 + 100) = 0.10$

$$G \text{ ratio} = 26 / (310 + 50 + 26 + 100) = 0.05$$

$$T \text{ ratio} = 100 / (310 + 50 + 26 + 100) = 0.21$$

55 At step 622 if either the reference experiments or the sample sequence failed to have sufficient intensity, the unknown base is assigned the code N (insufficient intensity) as shown at step 624.

At step 626 the wild-type and "other" ratios associated with the sample sequence are compared to statistical expressions. The statistical expressions include four predetermined standard deviation cutoffs, one associated with each base.

EP 0 717 113 A2

Thus, there is a standard deviation cutoff for each of the bases A, C, G, and T. The localized standard deviation cutoffs allow the unknown base to be called with higher precision because each standard deviation cutoff can be set to a different value. Suppose the standard deviation cutoffs are set as follows:

5 A standard deviation cutoff -> 4

C standard deviation cutoff -> 2

G standard deviation cutoff -> 8

10 T standard deviation cutoff -> 4

The wild-type base ratio associated with the sample is compared to a corresponding statistical expression:

15 WT ratio \geq WT mean - (WT std. dev. * WT base std. dev. cutoff)

where the WT base std. dev. cutoff is the standard deviation cutoff for the wild-type base. As the wild-type base is G, the above comparison solves to the following:

20 $0.05 \geq 0.71 - (0.050 * 8)$

0.05 \approx 0.31

which is not a true expression (0.05 is not greater than 0.31).

25 Each of the "other" ratios associated with the sample is compared to a corresponding statistical expression:

$$\text{Other ratio} > \text{Other mean} + (\text{Other std. dev.} * \text{Other base std. dev. cutoff})$$

where the Other base std. dev. cutoff is the standard deviation cutoff for the particular "other" base. Thus, the above
30 comparison solves to the following three expressions:

$$A \rightarrow 0.64 > 0.16 + (0.003 * 4)$$
 $0.64 > 0.17$

35 $C \rightarrow 0.10 > 0.03 + (0.002 * 2)$

$$0.10 > 0.03$$

40 $T \rightarrow 0.21 > 0.11 + (0.004 \cdot 4)$

$$0.21 > 0.13$$

which are all true expressions.

45 At step 628 if only the wild-type ratio of the sample sequence was greater than the statistical expression, the unknown base is assigned the code of the reference wild-type base as shown at step 630.

At step 632 if one or more of the "other" ratios of the sample sequence were greater than their respective statistical expressions, the unknown base is assigned an ambiguity code that indicates the unknown base may be any one of the complements of these bases, including the reference wild-type. In this example, the "other" ratios for A, C, and T were all greater than their corresponding statistical expression. Thus, the unknown base would be called the complements of these bases, represented by the subset T, G, and A. Thus, the unknown base would be assigned the code D (meaning "not C").

If none of the ratios are greater than their respective statistical expressions, the unknown base is assigned the code X (insufficient discrimination) as shown at step 636.

55 The statistical method provides accurate base calling because it utilizes statistical data from multiple reference experiments to call the unknown base. The statistical method has also been used to implement confidence estimates and calling of mixed sequences.

V. Pooling Processing

The present invention provides pooling processing which is a method of processing reference and sample nucleic acid sequences together to reduce variations across individual experiments. In the representative embodiment discussed herein, the reference and sample nucleic acid sequences are labeled with different fluorescent markers emitting light at different wavelengths. However, the nucleic acids may be labeled with other types of markers including distinguishable radioactive markers.

After the reference and sample nucleic acid sequences are labeled with different color fluorescent markers, the labeled reference and sample nucleic acid sequences are then combined and processed together. An apparatus for detecting targets labeled with different markers is provided in U.S. Application No. 08/195,889 and is hereby incorporated by reference for all purposes.

Fig. 13 illustrates the pooling processing of a reference and sample nucleic acid sequence. At step 702 a reference nucleic acid sequence is marked with a fluorescent dye, such as fluorescein. At step 704 a sample nucleic acid sequence is marked with a dye that, upon excitation, emits light of a different wavelength than that of the fluorescent dye of the reference sequence. For example, the sample nucleic acid sequence may be marked with rhodamine. Alternatively, the sample nucleic acid sequence may be marked by attaching biotin to the sample sequence which will subsequently bind to streptavidin labeled with phycoerythrin. Of course, either sequence may be marked with these or other dyes or other kinds of markers (e.g., radioactive) as long as the other sequence is marked with a marker that is distinguishable.

At step 706 the labeled reference sequence and the labeled sample sequence are combined. After this step, processing continues in the same manner as for only one labeled sequence. At step 708 the sequences are fragmented. The fragmented nucleic acid sequences are then hybridized on a chip containing probes as shown at step 710.

At step 712 a scanner generates image files that indicate the locations where the labeled nucleic acids bound to the chip. There is typically some overlap between the two signals. This is corrected for prior to further analysis, i.e., after correction, the data files correspond to "reference" and "sample." In general, the scanner generates an image file by focusing excitation light on the hybridized chip and detecting the fluorescent light that is emitted. The marker emitting the fluorescent light can be identified by the wavelength of the light. For example, the fluorescence peak of fluorescein is about 530 nm while that of a typical rhodamine dye is about 580 nm.

The scanner creates an image file for the data associated with each fluorescent marker, indicating the locations where the correspondingly labeled nucleic acid bound to the chip. Based upon an analysis of the fluorescence intensities and locations, it becomes possible to extract information such as the monomer sequence of DNA or RNA.

Pooling processing reduces variations across individual experiments because much of the test environment is common. Although pooling processing has been described as being used to improve the combined processing of reference and sample nucleic acid sequences, the process may also be used for two reference sequences, two sample sequences, or multiple sequences by utilizing multiple distinguishable markers.

Pooling processing may also be utilized with methods of the present invention of identifying mutations in a sample nucleic acid sequence. These methods are highly accurate in identifying single mutations, locating multiple mutations and removing false positives for mutations, where a false positive is a base that has erroneously been identified as a mutation. These methods utilize hybridization data from more than one base position to identify the likely position of mutations. The interrogation position on the probes is utilized to more accurately identify likely mutations which makes more efficient use of base calling methods. These methods may be advantageously combined with the base calling methods described herein to efficiently and accurately sequence a sample nucleic acid sequence.

As discussed earlier in reference to Fig. 8, the fluorescent intensities of cells near an interrogation position having a mutation are relatively dark which creates "dark regions" around the mutation. These lower fluorescent intensities result because the cells at interrogation positions near a mutation do not contain probes that are perfectly complementary to the sample sequence. Thus, the hybridization of these probes with the sample sequence is lower. The characteristics of these "dark regions" may be utilized to identify mutations and false positives.

For example, a sample sequence and a reference sequence were labeled with different fluorescent markers, in this case fluorescein and biotin/phycoerythrin. The sample and reference sequences are known and the sample sequence is identical to the reference sequence except for mutations at certain known positions. The sample and reference sequences were then processed together using the pooling processing described above and the sequences were hybridized to a chip including wild-type probes that are perfectly complementary to the reference sequence. The chip included 20-mer probes with the interrogation position of each probe being at the 12th base position in the probe.

Fig. 14A shows a graph of the scaled fluorescent intensities (photon counts) of the wild-type probes hybridizing with the sample and reference sequences. Along the bottom of the graph are numbers which represent wild-type cell positions on the chip. The photon counts of the probes in the wild-type cells are plotted on a logarithmic scale of 10^n . As shown, the photon counts range from 1 (representing a de minimus value) and 100,000. The photon counts for the probes in the wild-type cell numbered "45" is around 10,000.

At various wild-type cells, the photon count for the probes in the cells drops to 1 or lower. For example, the photon counts for wild-type cells numbered 11, 24, 39, etc. are 1. The low photon counts are due to the fact that there are no

probes in these cells. The cells are left "blank" in order to minimize diffraction edges and thus, the location of these blank cells is known. Consequently, the intermittent wild-type cells that have a photon count of 1 do not represent erroneous data.

As shown in Fig. 14A, the scaled photon counts for the wild-type probes hybridizing with the sample and reference sequences are almost the same except for two "bubbles." A bubble 730 has a top curve defined by the photon counts of the wild-type probes that hybridized with the reference sequence and a bottom curve defined by the photon counts of the wild-type probes that hybridized with the sample sequence. Following bubble 730, there is a section 732 where the photon counts for the wild-type probes hybridizing with the sample and reference sequences are almost the same. After section 732 is another bubble 734 which again has a top curve defined by the hybridization of the reference sequence and the bottom curve defined by the hybridization of the sample sequence. Another partial bubble is shown to the right of bubble 734.

Each bubble in Fig. 14A corresponds to a dark region surrounding a single mutation. Because the wild-type probes at and surrounding a mutant position in the sample sequence contain a single base mismatch with the sample sequence, the hybridization is relatively lower which results in lower photon counts. Much information about the sample sequence may be acquired by a detailed analysis of these bubble regions.

The width of the bubble indicates whether there is a false positive, a single mutation or a multiple mutation. If there is a single mutation, the width of the bubble should be approximately equal to the probe length. For example, Fig. 14A was produced utilizing 20-mer probes. Accordingly, bubbles 730 and 734 are approximately 20 wild-type cells wide indicating that the both these bubbles were produced by single mutations. The width of the dark region resulting from a single mutation is believed to be approximately equal to the probe length because each of the probes in this region have a single base mismatch with the sample sequence.

If the width of the bubble is substantially less than the probe length, the bubble may represent a false positive. For example, assume that at wild-type cell number 45 in Fig. 14A, the hybridization of the wild-type probe with the sample sequence was very low (e.g., around 1000 photon counts). A base calling algorithm that calls the bases according to the intensities among the cells at that position may indicate that there is a mutation at this position. However, the low photon counts may be due to dust on the chip and not due to lower hybridization. Since the width of this bubble would be 1, which is substantially lower than the probe width of 20, the lower photon count at wild-type cell 45 would not be due to a mutation (i.e., there is no dark region surrounding that position).

If the width of the bubble is substantially more than the probe length, the bubble may represent multiple mutations. In other words, the bubble may be produced by more than one overlapping dark region. The analysis of such a bubble will be discussed in more detail in reference to Fig. 14C.

Returning to Fig. 14A, each of bubbles 730 and 734 are approximately 20 bases wide indicating with a high degree of certainty that each of the bubbles represent a single mutation. Furthermore, the bubbles may be analyzed to determine the probable location of the mutations within the bubbles. As mentioned earlier, the 20-mer probes on the chip had an interrogation position at the 12th base position in the probe. Thus, the base at the 12th base position is the base that varies among the related WT-, A-, C-, G- and T-cells. Accordingly, the mutation should be located at the 12th position in the bubble.

The actual mutation in bubble 730 occurs at the 12th position (from the left). Additionally, the actual mutation in bubble 734 occurs at the 12th position (from the left). Thus, as the graph shows, there are 11 bases to the left of each mutation and 8 bases to the right of each mutation. By utilizing the location of the interrogation position within the probes, the present invention can help to identify the probable location of a mutation within a dark region or bubble.

Additionally, because this method identifies specific locations that may have a mutation, more efficient base calling may be achieved. For example, an analysis of bubble 730 indicates that there is likely to be a single mutation around wild-type cell 15. Typically, most errors in base calling occur in the dark regions surrounding a mutation. Many false positives in this dark zone can now be eliminated because they are incompatible with the bubble size (which indicates single mutation, for example). Also, by identifying clearly a "mismatch zone," we can now apply algorithms that factor in the effect of a mismatch or multiple mismatches.

Additionally, the shape of the bubble may indicate what mutation has occurred. Fig. 14B shows a hypothetical graph of the fluorescent intensities vs. cell locations for wild-type probes hybridizing with two sample sequences and one reference sequence. A C-A mismatch will be more destabilizing to probe hybridization than a U-G mismatch. As shown, the more destabilizing C-A mismatch results in a larger volume bubble. The shape of the bubble may be utilized to identify the particular mutation by pattern matching bubbles stored in a library.

Fig. 14C shows a graph of the fluorescent intensities (photon counts) of the wild-type probes hybridizing with the sample and reference sequences. A single bubble 750 is flanked on either side by regions 752 and 754 which do not contain a mutation. The graph was produced from a chip containing 20-mer probes with an interrogation position at base 12 on the probes.

As shown, bubble 750 is 27 bases wide indicating that the bubble was produced from the dark regions surrounding more than one mutation as 27 is greater than 20 or the length of the probes. In addition to providing information that there are multiple mutations, analysis of the bubble indicates the probable position of two of the mutations. Because the

interrogation position is at base 12 in the 20-mer probes, one of the mutations should be around 12 bases from the left end of the bubble while another mutations should be around 8 bases from the right end of the bubble. And in fact, there is a mutation of T to C at wild-type cell 62 which is 12 bases from the left of the bubble. Additionally, there is a mutation of A to G at wild-type cell 69 which is 8 bases from the right of the bubble.

5 The third and last mutation within bubble 750 may be identified by performing base calling methods within the bubble. Alternatively, the mutation may be identified by pattern matching bubbles from a library that indicate not only the number of mutations but also the specific location and type of mutation.

Fig. 15 illustrates the high level flow of one embodiment of the present invention that uses the hybridization data from more than one base position to identify mutations in a sample nucleic acid sequence. After probe intensities from the hybridization of wild-type probes with a sample and reference sequence are measured, the system identifies a bubble region at step 780. Bubble regions are identified as regions where the hybridization of the wild-type probes to the sample and reference sequence differ significantly. Additionally, the reference sequence should hybridize more strongly with the wild-type probes since the wild-type probes will be perfectly complementary to the reference sequence.

At step 782, the system compares the base width of the bubble to the probe length. If the bubble width is substantially less than the probe length, the bubble does not represent a mutation at step 784. The determination of how much less the bubble width may vary according to experiment conditions.

At step 786, the system compares the base width of the bubble to the probe length to determine if they are approximately equal. If the bubble width is approximately equal to the probe length, the bubble represents a single base mutation at step 788. Again, the determination of how close the bubble width should be to the probe length may vary according to experiment conditions.

If the bubble width is substantially more than the probe length, the bubble represents multiple mutations at step 790. The system performs base calling at likely locations of mutations at step 792. The likely locations of mutations are determined by both the width of the bubble and the location of the interrogation position on the probes. Additionally, the system may analyze the pattern of the bubble to determine the specific mutations and their positions by analyzing the pattern of the bubble. The base calling method with the present invention may be the intensity ratio method, reference method, statistical method, or any other method.

At step 794, the system produces confidences that the mutations are identified correctly. Each confidence is determined by how closely the experimental data matched the data expected for the mutation that was called. For example, if the bubble width was exactly the same as the probe length and the base calling method identified a mutation at the interrogation position in the probes, there is a very high likelihood or probability that the mutation was identified correctly. The confidence may also be produced according to how closely the bubble pattern matched the pattern for that mutation or mutations in the library of patterns.

Although in a preferred embodiment, this method of identifying mutations in a sample nucleic acid sequence is utilized in conjunction with pooling processing in order to reduce variations, the method may be utilized without pooling processing. For example, the method may be utilized effectively where the variations between separate experiments is minimized or the data is adjusted accordingly. Therefore, this method is not limited to the embodiment discussed above.

The present invention provides methods of accurately identifying single mutations, locating multiple mutations and removing false positives for mutations. These methods are advantageously performed with pooling processing and utilize hybridization data from more than one base position to identify the likely position of mutations. The interrogation position on the probes is also utilized to more accurately identify the likely position of mutations which makes more efficient use of base calling methods.

VI. Comparative Analysis (ViewSeq™)

45 The present invention provides a method of comparative analysis and visualization of multiple experiments. The method allows the intensity ratio, reference, and statistical methods to be run on multiple datafiles simultaneously. This permits different experimental conditions, sample preparations, and analysis parameters to be compared in terms of their effects on sequence calling. The method also provides verification and editing functions, which are essential to reading sequences, as well as navigation and analysis tools.

Fig. 16 illustrates the main screen and the associated pull down menus for comparative analysis and visualization of multiple experiments (SEQ ID NO:8 and SEQ ID NO:9). The windows shown are from an appropriately programmed Sun Workstation. However, the comparative analysis software may also be implemented on or ported to a personal computer, including IBM PCs and compatibles, or other workstation environments. A window 802 is shown having pull down menus for the following functions: File 804, Edit 806, View 808, Highlight 810, and Help 812.

55 The main section of the window is divided into a reference sequence area 814 and a sample sequence area 816. The reference sequence area is where known sequences are displayed and is divided into a reference name subarea 818 and reference base subarea 820. The reference name subarea is shown with the filenames that contain the reference sequences. The chip wild-type is identified by the filename with the extension ".wt#" where the # indicates a unit on the chip. The reference base subarea contains the bases of the reference sequences. A capital C 822 is displayed to the

right of the reference sequence that is the chip wild-type for the current analysis. Although the chip wild-type sequence has associated fluorescence intensities, the other reference sequences shown below the chip wild-type may be known sequences that have not been tiled on the chip. These may or may not have associated fluorescence intensities. The reference sequences other than the chip wild-type are used for sequence comparisons and may be in the form of simple ASCII text files.

Sample sequence area 816 is where sample or unknown experimental sequences are displayed for comparison with the reference sequences. The sample sequence area is divided into a sample name subarea 824 and sample base subarea 826. The sample name subarea is shown with filenames that contain the sample sequences. The filename extensions indicate the method used to call the sample sequence where ".cq#" denotes the intensity ratio method, ".rq#" denotes the reference method, and ".sq#" denotes the statistical method (# indicates the unit on the chip). The sample base subarea contains the bases of the sample sequences. The bases of the sample sequences are identified by the codes previously set forth which, for the most part, conform to the IUPAC standard.

Window 802 also contains a message panel 828. When the user selects a base with an input device in the reference or sample base subarea, the base becomes highlighted and the pathname of the file containing the base is displayed in the message panel. The base's position in the nucleic acid sequence is also displayed in the message panel.

In pull down menu File 804, the user is able to load files of experimental sequences that have been tiled and scanned on a chip. There is a chip wild-type associated with each experimental sequence. The chip wild-type associated with the first experimental sequence loaded is read and shown as the chip wild-type in reference sequence area 814. The user is also able to load files of known nucleic acid sequences as reference sequences for comparison purposes. As before, these known reference sequences may or may not have associated probe intensity data. Additionally, in this menu the user is able to save sequences that are selected on the screen into a project file that can be loaded in at a later time. The project file also contains any linkage of the sequences, where sequences are linked for comparison purposes. Sequences to be saved, both reference and sample, are chosen by selecting the sequence filename with an input device in the reference or sample name subareas.

In pull down menu Edit 806, the user is able to link together sequences in the reference and sample sequence areas. After the user has selected one reference and one or more sample sequences, the sample sequences can be linked to the reference sequence by selecting an entry in the pull down menu. Once the sequences are linked, a link number 830 is displayed next to each of sequences of related interest. Each group of linked sequences is associated with a unique link number, so the user can easily identify which sequences are linked together. Linking sequences permits the user to more easily compare the linked sequences. The user is also able to remove and display links from this menu.

In pull down menu View 808, the user is able to display intensity graphs for selected bases. Once a base is selected in the reference or sample base subareas, the user may request an intensity graph showing the hybridized probe intensities of the selected base and a delineated neighborhood of bases near the selected base. Intensity graphs may be displayed for one or multiple selected bases. The user is also able to prepare comment files and reports in this menu.

Fig. 17 illustrates an intensity graph window for a selected base at position 120 (SEQ ID NO:30 and SEQ ID NO:31). The filename containing the sequence data is displayed at 904. The graph shows the intensities for each of the hybridized probes associated with a base. Each grouping of four vertical bars on the graph, which are labeled as "a", "c", "g", and "t" on line 906, shows the background subtracted intensities of probes having the indicated substitution base. In one embodiment, the called bases are shown in red. The wild-type base is shown at line 908, the called base is shown at line 910, and the base position is shown at line 912. In Fig. 17, the base selected is at position 120, as shown by arrow 914. The wild-type base at this position is T; however, the called base is M which means the base is either A or C (amino). The user is able to use intensity graphs to visually compare the intensities of each of the possible calls.

Fig. 18 illustrates multiple intensity graph windows for selected bases (SEQ ID NO:32, SEQ ID NO:33, SEQ ID NO:34, and SEQ ID NO:35). There are three intensity graph windows 1002, 1004, and 1006 as shown. Each window may be associated with a different experiment, where the sequence analyzed in the experiment may be either a reference (if it has associated probe intensity data as in the chip wild-type) or a sample sequence. The windows are aligned and a rectangular box 1008 shows the selected bases' position in each of the sequences (position 162 in Fig. 18). The rectangular box aids the user in identifying the selected bases.

Referring again to Fig. 16, in pull down menu Highlight 810, the user is able to compare the sequences of references and samples. At least four comparisons are available to the user, including the following: sample sequences to the chip wild-type sequence, sample sequences to any reference sequences, sample sequences to any linked reference sequences, and reference sequences to the chip wild-type sequence. For example, after the user has linked a reference and sample sequence, the user can compare the bases in the linked sequences. Bases in the sample sequence that are different from the reference sequence will then be indicated on the display device to the user (e.g., base is shown in a different color). In another example, the user is able to perform a comparison that will help identify sample sequences. After a sample is linked to multiple reference sequences, each base in the sample sequence that does not match the wild-type sequence is checked to see if it matches one of the linked reference sequences. The bases that match a linked

reference sequence will then be indicated on the display device to the user. The user may then more easily identify the sample sequence as being one of the reference sequences.

In pull down menu Help 812, the user is able to get information and instructions regarding the comparative analysis program, the calling methods, and the IUPAC definitions used in the program.

Fig. 19 illustrates the intensity ratio method correctly calling a mutation in solutions with varying concentrations (SEQ ID NO:10, SEQ ID NO:11, SEQ ID NO:12, SEQ ID NO:13, SEQ ID NO:14, SEQ ID NO:15, SEQ ID NO:16, SEQ ID NO:17, and SEQ ID NO:18). A window 1102 is shown with a chip wild-type 1104 and a mutant sequence 1106. The mutant sequence differs from the chip wild-type at the position indicated by the rectangular box 1108. The chip wild-type and mutant sequences are a region of HIV Pol Gene spanning mutations occurring in AZT drug therapy.

There are seven sample sequences that are called using the intensity ratio method. The sample sequences are actually solutions of different proportions of the chip wild-type sequence and the mutant sequence. Thus, there are sample solutions 1110, 1112, 1114, 1116, 1118, 1120, and 1122. The solutions are 15-mer tilings across the chip wild-type with increased percentages of the mutant sequence from 0 to 100% by weight. The following shows the proportions of the sample solutions:

Sample Solution	Chip Wild-Type:Mutant
1110	100:0
1112	90:10
1114	75:25
1116	50:50
1118	25:75
1120	10:90
1122	0:100

For example, sample solution 1114 contains 75% chip wild-type sequence and 25% mutant sequence.

Now referring to the bases called in rectangular box 1108 for the sample solutions, the intensity ratio method correctly calls sample solution 1110 as having a base A as in the chip-wild type sequence. This is correct because sample solution 1110 is 100% chip wild-type sequence. The intensity ratio method also calls sample solution 1112 as having a base A because the sample solution is 90% chip wild-type sequence.

The intensity ratio method calls the identified base in sample solutions 1114 and 1116 as being an R, which is an ambiguity IUPAC code denoting A or G (purine). This also a correct base call because the sample solutions have from 75% to 50% chip-wild type sequence and from 25% to 50% mutation sequence. Thus, the intensity ratio method correctly calls the base in this transition state.

Sample solutions 1118, 1120, and 1122 are called by the intensity ratio method as having a mutation base G at the specified location. This is a correct base call because the sample solutions primarily consist of the mutation sequence (75%, 90%, and 100% respectively). Again, the intensity ratio method correctly called the bases.

These experiments also show that the base calling methods of the present invention may also be used for solutions of more than one nucleic acid sequence.

Fig. 20 illustrates the reference method correctly calling a mutant base where the intensity ratio method incorrectly called the mutant base (SEQ ID NO:36, SEQ ID NO:37, SEQ ID NO:38, and SEQ ID NO:39). There are three intensity graph windows 1202, 1204, and 1206 as shown. The windows are aligned and a rectangular box 1208 outlines the bases of interest. Window 1202 shows a sample sequence called using the intensity ratio method. However, the base in the rectangular box 1208 was incorrectly called base C, as there is actually a base A at that position. The intensity ratio method incorrectly called the base as C because the probe intensity associated with base C is much higher than the other probe intensities.

Window 1204 shows a reference sequence called using the intensity ratio method. As the reference sequence is known, it is not necessary to know the method used to call the reference sequence. However, it is important to have probe intensities for a reference sequence to use the reference method. The reference sequence is called a base C at the position indicated by the rectangular box.

Window 1206 shows the sample sequence called using the reference method. The reference method correctly calls the specified base as being base A. Thus, for some cases the reference method is preferable to the intensity ratio method because it compares probe intensities of a sample sequence to probe intensities of a reference sequence.

VII. ExamplesExample 1

The intensity ratio method was used in sequence analysis of various polymorphic HIV-1 clones using a protease chip. Single stranded DNA of a 382 nt region was used with 4 different clones (HXB2, SF2, NY5, pPol4mut18). Results were compared to results from an ABI sequencer. The results are illustrated below:

	ABI		Protease Chip	
	Sense	Antisense	Sense	Antisense
No call	0	4	9	4
Ambiguous	6	14	17	8
Wrong call	2	3	3	1
TOTAL	8	21	29	13
SUMMARY				
ABI (sense) - 99.5%				
Chip (sense) - 98.1%				
ABI (antisense) - 98.6%				
Chip (antisense) - 99.1%				

Example 2

HIV protease genotyping was performed using the described chips and CallSeq™ intensity ratio calculations. Samples were evaluated from AIDS patients before and after ddI treatment. Results were confirmed with ABI sequencing.

Fig. 21 illustrates the output of the ViewSeq™ program with four pretreatment samples and four posttreatment samples (SEQ ID NO:22, SEQ ID NO:23, SEQ ID NO:24, SEQ ID NO:25, SEQ ID NO:26, and SEQ ID NO:27). Note the base change at position 207 where a mutation has arisen. Even adjacent two additional mutations (gt), the "a" mutation has been properly detected.

The above description is illustrative and not restrictive. Many variations of the invention will become apparent to those of skill in the art upon review of this disclosure. Merely by way of example, while the invention is illustrated with particular reference to the evaluation of DNA (natural or unnatural), the methods can be used in the analysis from chips with other materials synthesized thereon, such as RNA. The scope of the invention should, therefore, be determined not with reference to the above description, but instead should be determined with reference to the appended claims along with their full scope of equivalents.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT:

- (A) NAME: Affymax Technologies N.V.
- (B) STREET: De Ruyderkade 62
- (C) CITY: Curacao
- (E) COUNTRY: Netherlands Antilles
- (F) POSTAL CODE (ZIP): none

(ii) TITLE OF INVENTION: Computer-Aided Visualization and Analysis System for Sequence Evaluation

(iii) NUMBER OF SEQUENCES: 39

(iv) COMPUTER READABLE FORM:

- (A) MEDIUM TYPE: Floppy disk
- (B) COMPUTER: IBM PC compatible
- (C) OPERATING SYSTEM: PC-DOS/MS-DOS
- (D) SOFTWARE: PatentIn Release #1.0, Version #1.25 (EPO)

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 15 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

ATGTGGACAG TTGTA

15

(2) INFORMATION FOR SEQ ID NO:2:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 15 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

ATGTGGATAG TTGTA

15

(2) INFORMATION FOR SEQ ID NO:3:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 15 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

ATGTGGAKAG TTGTA

15

(2) INFORMATION FOR SEQ ID NO:4:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

AAAACTGAAA A

11

(2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

AAAACCGAAA A

11

(2) INFORMATION FOR SEQ ID NO:6:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

AAACCCAATC CACATCA

17

(2) INFORMATION FOR SEQ ID NO:7:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

AAACCCAGTC CACATCA

17

(2) INFORMATION FOR SEQ ID NO:8:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 31 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

GGGGAAGCAG ATTTGGGTAC CACCCAAGTA T

31

(2) INFORMATION FOR SEQ ID NO:9:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 31 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

GGGGAAGCAG ATTTGAAMAC CACCCAAGTA T

31

(2) INFORMATION FOR SEQ ID NO:10:

- 5 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear
- 10 (ii) MOLECULE TYPE: DNA (oligonucleotide)

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GCATTAGTAG AGATATGTAC AGAAATGGAA AAGGAAGGGA AAATTTCAAA
AATTGGGCC

59

20

(2) INFORMATION FOR SEQ ID NO:11:

- 25 (i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
30 (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA (oligonucleotide)

35

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GCATTAGTAG AAATTTGTAC AGAGATGGAA AAGGAAGGGA AAATTTCAAA
AATTGGGCC

59

40

45

50

55

(2) INFORMATION FOR SEQ ID NO:12:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 59 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GCATTAGTAG AGATATGGAG AGRARDGGRA AXXXAAGGGA AAATTNNNAA
 AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:13:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 59 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GCATTAGTAG AGATATGKAS AGRARDGGRA AXXXAAGGGA AAKTNNNAA
 AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:14:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 59 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GCATTAGTAG AGATATGKAS AGRRRDGGRA AXXXAAGGGA AAADTYNNAA
AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:15:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 59 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GCATTAGTAG AGATATGTAS AGRRADGGAA AXGGAAGGGA AAATTNNNNA
AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:16:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 59 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GCATTAGTAG AGATATGTAC AGRGAGGGAA AXGGAAGGGA AAATTNNNNA
 AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:17:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 59 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GCATTAGTAG AGATATGTAS AGRGAGGGAA AXGGAAGGGA AAATTNNNNA
 AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 59 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

GCATTAGTAG GAGGNNNGAC AGGGRKGGAA AXXMAAGGGA AAAKTNNNAA
AATTGGGCC

59

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 160 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACTTAA
ACGGTCCTTT

60

TACCTTTGGT TTTTACTATC CCCCTTAACC TCCAAAATAG TTTCATTCTG
TCATGCTAGT

120

CTATGGACAT CTTTAGACAC CTGTATTTTCG ATATCCATGT

160

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

NNGAGATANN NTATGTCCTC GTCYACTATG TNANNNNNNNN NNNNNNNNAA 60
 ACGGTCCTNN
 NNNNNNNNNN NNNNNNNNNN CNNCNTAACC TCCAAAATAN NNNNNNTCTN 120
 NNNNANNNNT
 CTANNNGNAG NNNNAGANAR NCCNNNNNNN NNATNCATGT 160

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATNNN NNNNACTTAA 60
 ACGGTCCTTT
 TACCTTTGGT TTTTACTATC CCCCTTAACC TCCAAAATAG TTTCATTCTG 120
 NCATANNAGT
 CTATGNGNNG NNNTAGACAG NCCNNNNNTCG ATATCCATGT 160

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACTTAA	
ACGGTCCTTT	60
TACCTTTGGT TTTTACTATC CNNCTTAACC TCCAAAATAG TTTCATTCTG	
TCATACTAGT	120
CTATGGGTAG CTTTAGACCN CCGTATTTTCG ATATCCATGT	160

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACTTAA	
ACGGTCCTTT	60
TACCTTTGGT TTTTACTATC CCNCTTAACC TCCAAAATAG TTTCATTCTG	
TCATACTAGT	120
CTATGGGTAG CTTTAGACCC CCGTATTTTCG ATATCCATGT	160

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

NCGGGATANT NTATGTCCTC GTCYACTATG TCANNNNNCN NNCNNNNCAA	
ACGGTCCNCC	60
NNNNNCNNNN NNCNNCYANG AANCYCAACC TCCAAAATAN NNNNNNTCTN	
NNNNANNNCN	120
CTNNNNNNAG NGNNAGACAC CTGTATNNNN NTATNCAYGT	160

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 160 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

TCGRGATAAT CTATGTCCTC GTCTACTATG TCATAATCCN NNCNNCTCAA	
ACGGTCCCTYC	60
CNNNNYTGGT TNYTACTATC CCCCTTAACC TCCAAAATAG TTTCATTCTG	
NCATACNNST	120
CTANNNNNAG NGTTAGACAC CTGTATTTTCG ATATCCATGT	160

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 160 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCCN NCCTACTCAA
ACGGTCCTTC 60

TACCTTTGGT TTTTACTATC CMCCTTAACC TCCAAAATAG TTTCATTCTG
TCATACTAGT 120

CTATGAGTAG CTTTAGACAC CTGTATTTTCG ATATCCATGT 160

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 160 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

TCGAGATAAT CTATGTCCTC GTCTACTATG TCATAATCTT CTTTACYCAA
ACGGTCCTXC 60

TACCTTTGGT TTTTACTATC CCMCTTAACC TCCAAAATAG TTTCATTCTG
TCATACTAGT 120

CTATGAGTAG CTTTAGACAC CTGTATTTTCG ATATCCATGT 160

(2) INFORMATION FOR SEQ ID NO:28:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

AAACCCAATC CACATCM

17

(2) INFORMATION FOR SEQ ID NO:29:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 17 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

MMACNCANNC CACANNM

17

(2) INFORMATION FOR SEQ ID NO:30:

- 5 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- 10 (ii) MOLECULE TYPE: DNA (oligonucleotide)

15 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

TTGGGTACCA C 11

20

(2) INFORMATION FOR SEQ ID NO:31:

- 25 (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear
- 30 (ii) MOLECULE TYPE: DNA (oligonucleotide)

35 (xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

TTGAAMACCA C 11

40

45

50

55

(2) INFORMATION FOR SEQ ID NO:32:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

ACAGAAATGG A

11

(2) INFORMATION FOR SEQ ID NO:33:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

AGAGRATDGG R

11

(2) INFORMATION FOR SEQ ID NO:34:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

ASAGRRADGG A

11

(2) INFORMATION FOR SEQ ID NO:35:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 11 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: single
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

ACAGGGRRGG A

11

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 11 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

CTGGGGGGTA T

11

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 11 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

CTGGCCSGTG T

11

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 11 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CTGGGCGGTA T

11

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 11 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (oligonucleotide)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

CTGGCACGTG T

11

Claims

1. In a computer system, a method of identifying an unknown base in a sample nucleic acid sequence, said method comprising the steps of:
 - inputting a plurality of probe intensities, each of said probe intensities being associated with a nucleic acid probe;
 - said computer system comparing said plurality of probe intensities wherein each of said plurality of probe intensities is substantially proportional to said associated nucleic acid probe hybridizing with at least one nucleic acid sequence, said at least one nucleic acid sequence including said sample sequence; and
 - calling said unknown base according to results of said comparing step.
2. In a computer system, a method of identifying an unknown base in a sample nucleic acid sequence, said method comprising the steps of:
 - inputting a plurality of probe intensities, each of said probe intensities being associated with a nucleic acid probe;
 - said computer system comparing said plurality of probe intensities wherein each of said plurality of probe intensities is substantially proportional to said associated nucleic acid probe hybridizing with said sample sequence; and
 - calling said unknown base according to results of said comparing step.
3. The method of claim 2, wherein said comparing step includes the step of said computer system calculating a ratio of a higher probe intensity to a lower probe intensity.
4. The method of claim 3, wherein said calling step includes the step of calling said unknown base according to said probe associated with said higher probe intensity if said ratio is greater than a predetermined ratio value.
5. The method of claim 4, wherein said predetermined ratio value is approximately 1.2.
6. In a computer system, a method of identifying an unknown base in a sample nucleic acid sequence, said method comprising the steps of:
 - inputting a first set of probe intensities, each of said probe intensities in said first set being associated with a nucleic acid probe and substantially proportional to said associated nucleic acid probe hybridizing with a reference nucleic acid sequence;
 - inputting a second set of probe intensities, each of said probe intensities in said second set being associated with a nucleic acid probe and substantially proportional to said associated nucleic acid probe hybridizing with said sample sequence;

said computer system comparing at least one of said probe intensities in said first set and at least one of said probe intensities in said second set; and
calling said unknown base according to results of said comparing step.

- 5 7. The method of claim 6, wherein said comparing step includes the steps of:
calculating first ratios of a wild-type probe intensity to each probe intensity of a probe hybridizing with said reference sequence, wherein said wild-type probe intensity is associated with a wild-type probe; and
calculating second ratios of the highest probe intensity of a probe hybridizing with said sample sequence to each probe intensity of a probe hybridizing with said sample sequence.
- 10 8. The method of claim 7, wherein said comparing step further includes the step of calculating third ratios of said first ratios to said second ratios.
- 15 9. The method of claim 8, wherein said calling step includes the step of calling said unknown base according to said probe associated with a highest third ratio.
10. The method of claim 6, wherein said comparing step includes the step of calculating a ratio of a highest probe intensity in said first set to a highest intensity in said second set.
- 20 11. The method of claim 10, wherein said comparing step further includes the step of comparing said ratio of neighboring nucleic acid probes.
12. In a computer system, a method of identifying an unknown base in a sample nucleic acid sequence, said method comprising the steps of:
25 inputting statistics about a plurality of experiments, each of said experiments producing probe intensities each being associated with a nucleic acid probe and substantially proportional to said associated nucleic acid probe hybridizing with a reference nucleic acid sequence;
inputting a plurality of probe intensities, each of said plurality of probe intensities being associated with a nucleic acid probe and substantially proportional to said associated nucleic acid probe hybridizing with said sample
30 sequence;
said computer system comparing at least one of said plurality of probe intensities with said statistics; and
calling said unknown base according to results of said comparing step.
13. The method of claim 12, further comprising the step of calculating said statistics.
- 35 14. The method of claim 12, wherein said statistics include a mean and standard deviation.
15. A method of processing first and second nucleic acid sequences, comprising the steps of:
providing a plurality of nucleic acid probes;
40 labeling said first nucleic acid sequence with a first marker;
labeling said second nucleic acid sequence with a second marker; and
hybridizing said first and second labeled nucleic acid sequences at the same time.
16. The method of claim 15, wherein said plurality of nucleic acid probes are on a chip.
- 45 17. The method of claim 15, further comprising the step of fragmenting said first and second nucleic acid sequences at the same time.
18. The method of claim 15, further comprising the step of scanning for said first and second markers on said chip, said first and second labeled nucleic acid sequences being on said chip.
- 50 19. The method of claim 15, wherein said first and second markers are fluorescent markers that emit light at different wavelengths upon excitation.
- 55 20. In a computer system, a method of identifying mutations in a sample nucleic acid sequence, said method comprising the steps of:
inputting a first set of probe intensities, each of said probe intensities in said first set being associated with a nucleic acid probe and substantially proportional to said associated nucleic acid probe hybridizing with a reference nucleic acid sequence;

inputting a second set of probe intensities, each of said probe intensities in said second set being associated with a nucleic acid probe and substantially proportional to said associated nucleic acid probe hybridizing with said sample sequence;

said computer system comparing probe intensities in said first set and probe intensities in said second set to select hybridization regions where said probe intensities in said first set and said probe intensities in said second set differ; and

identifying mutations according to characteristics of said selected regions.

21. The method of claim 20, wherein said selected regions are determined by comparing probe intensities of wild-type probes.

22. The method of claim 21, wherein said wild-type probes are complementary to a portion of said reference sequence.

23. The method of claim 21, wherein said identifying step further includes the steps of:

analyzing a size of a selected region;

identifying a likely position of a mutation in said selected region according to an interrogation position of said nucleic acid probes; and

performing base calling at said likely position.

24. In a computer system, a method of analyzing a plurality of sequences of bases, said plurality of sequences including at least one reference sequence and at least one sample sequence, the method comprising the steps of:

displaying said at least one reference sequence in a first area on a display device; and

displaying said at least one sample sequence in a second area on said display device;

whereby a user is capable of visually comparing said plurality of sequences.

25. The method of claim 24, wherein said plurality of sequences are monomer strands of DNA or RNA.

26. The method of claim 24, wherein said at least one reference sequence includes a chip wild-type that has been tiled on a chip.

27. The method of claim 26, wherein said chip wild-type sequence is displayed as a first sequence in said first area.

28. The method of claim 26, further comprising the step of displaying a label in said first area to identify said chip wild-type sequence.

29. The method of claim 24, wherein said at least one sample sequence has been hybridized on a chip.

30. The method of claim 24, further comprising the step of indicating bases that differ among a plurality of user selected sequences.

31. The method of claim 24, further comprising the steps of:

displaying a name associated with each of said at least one reference sequence in said first area; and

displaying a name associated with each of said at least one sample sequence in said second area.

32. The method of claim 24, further comprising the step of linking at least one reference sequence in said first area with at least one sample sequence in said second area.

33. The method of claim 32, further comprising the step of indicating on said display device which sequences are linked.

34. The method of claim 24, further comprising the step of indicating bases of said at least one sample sequence that are not equal to a corresponding base in said at least one reference sequence.

35. The method of claim 24, wherein said at least one reference sequence and said at least one sample sequence are aligned on said display device. hybridization with said probes.

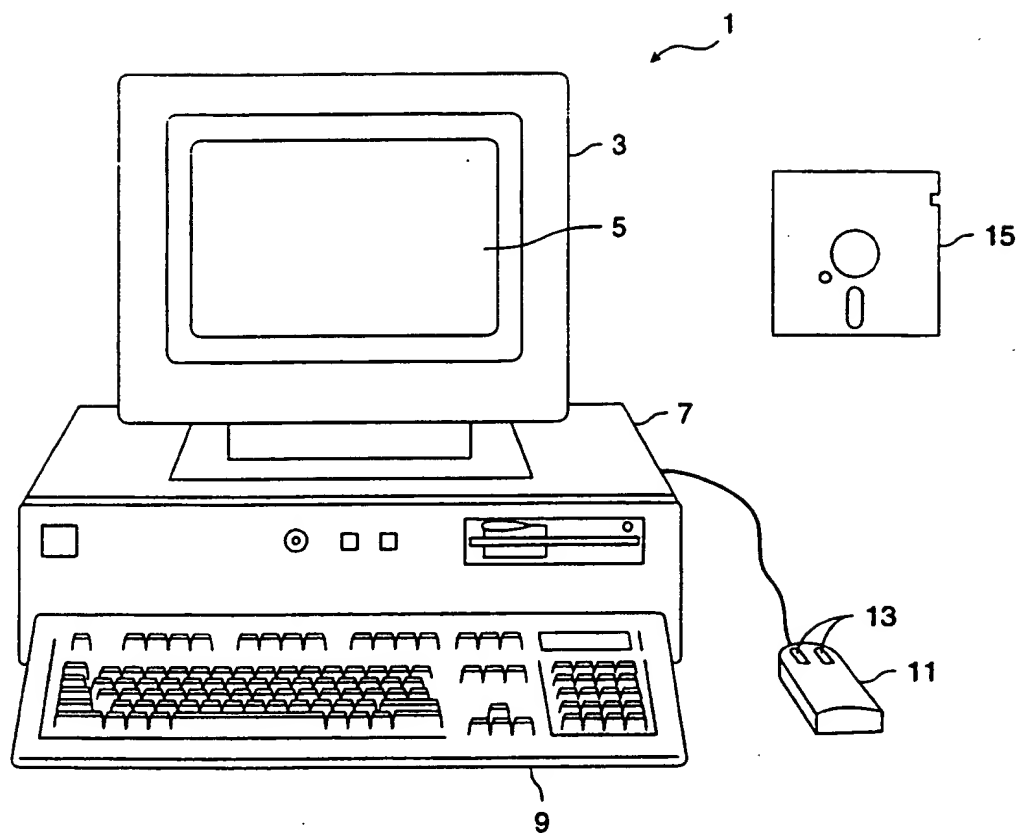


FIG. 1

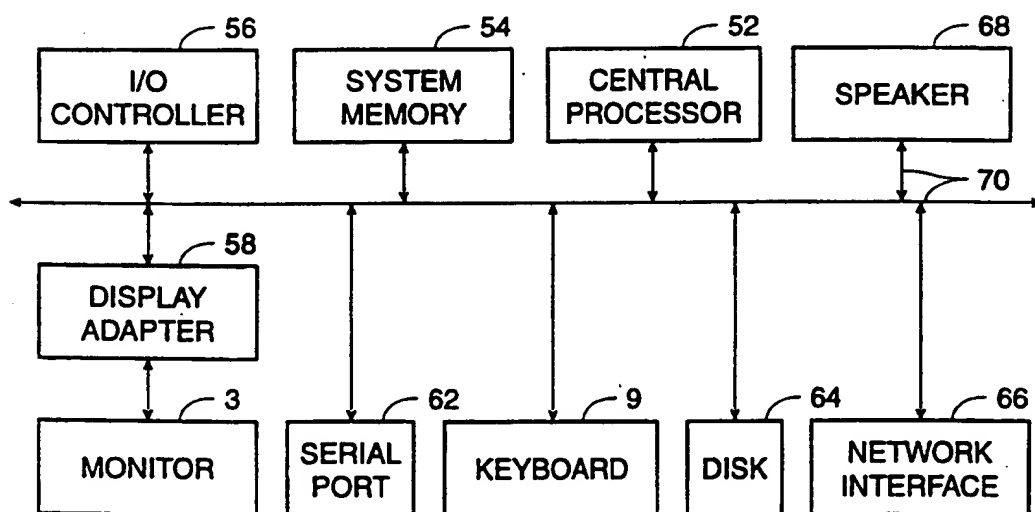


FIG. 2

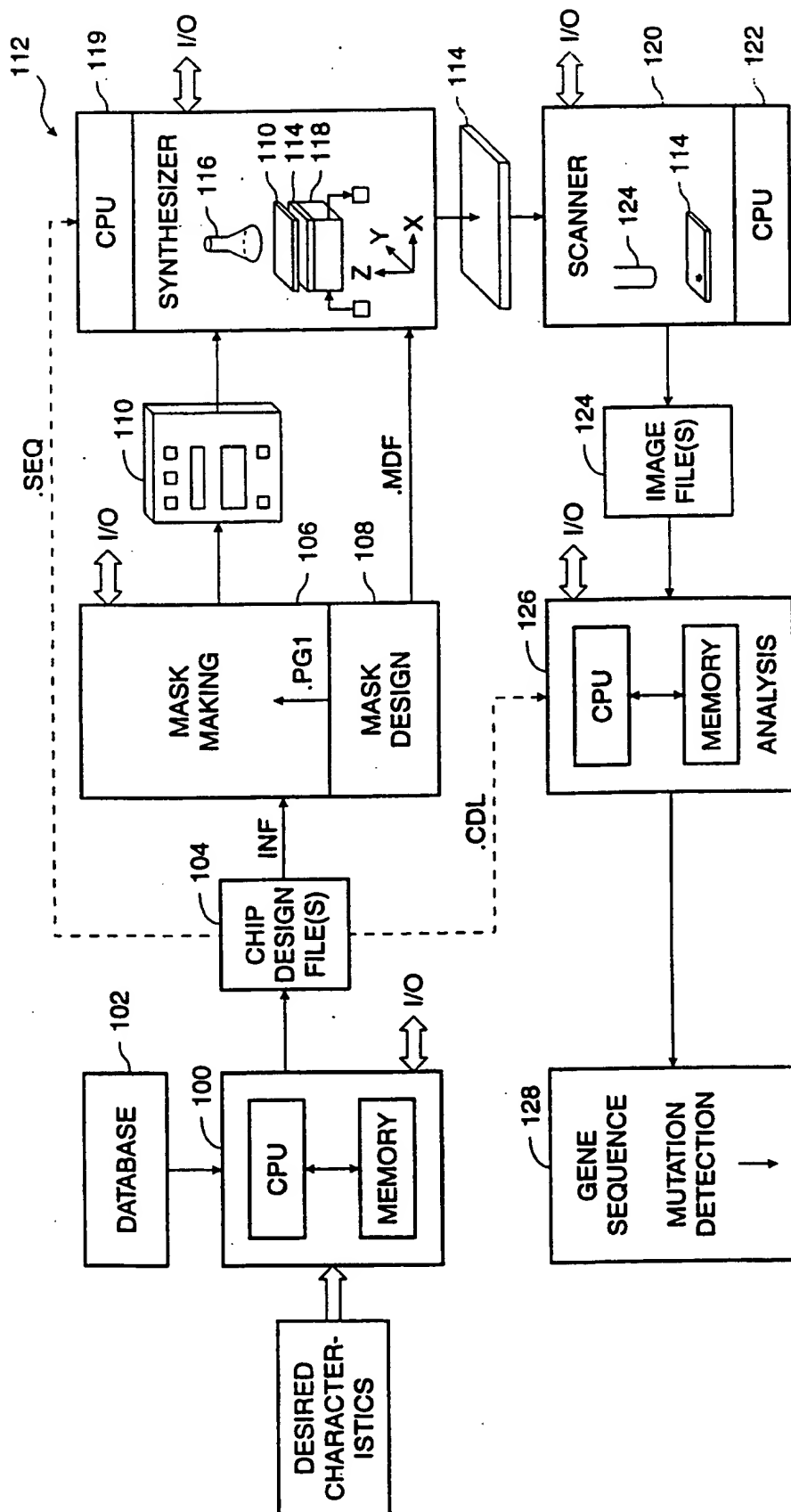


FIG. 3

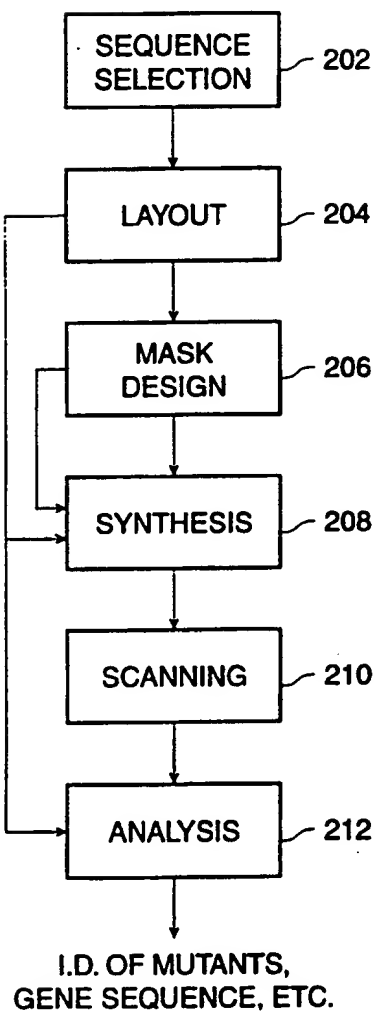


FIG. 4

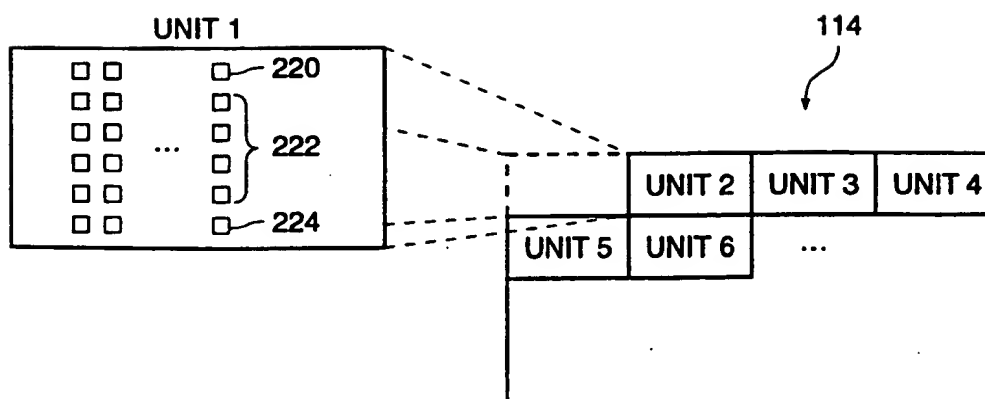


FIG. 5

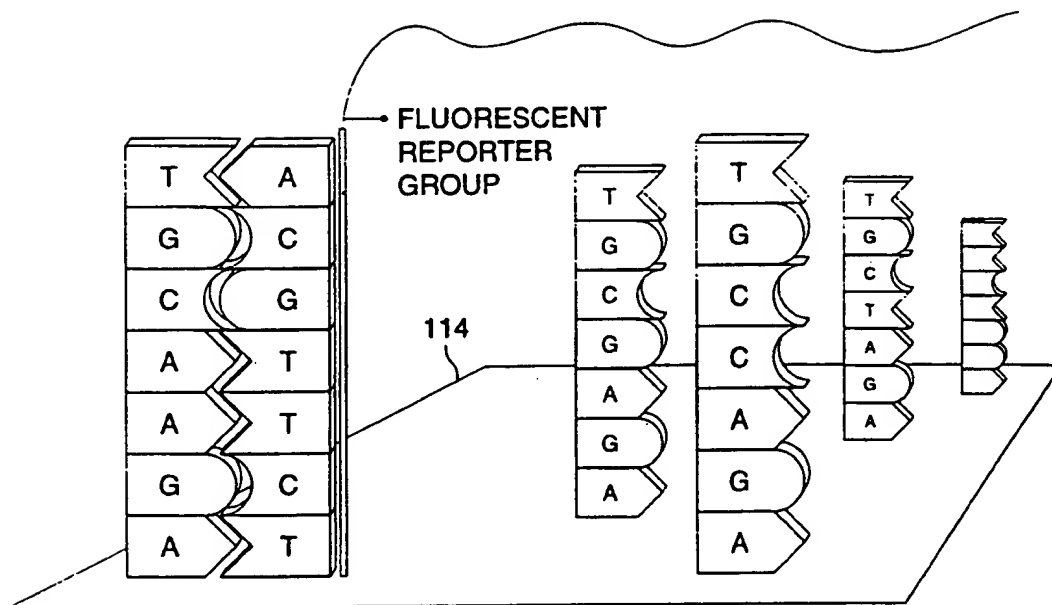


FIG. 6

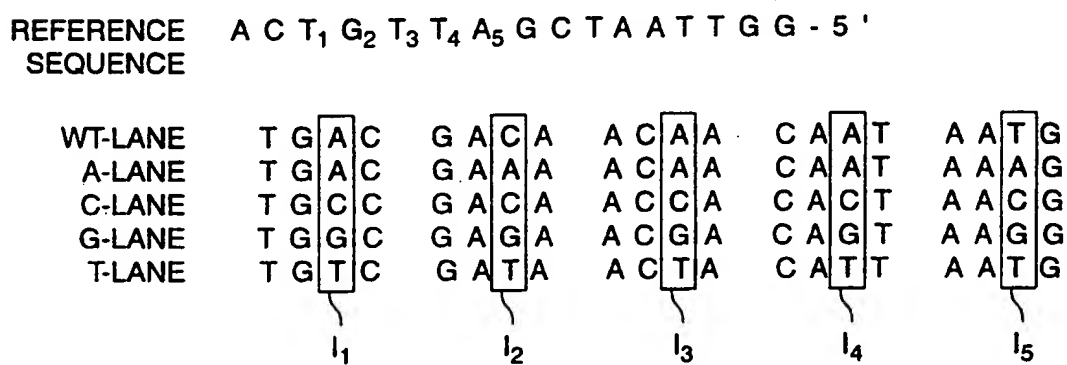


FIG. 7

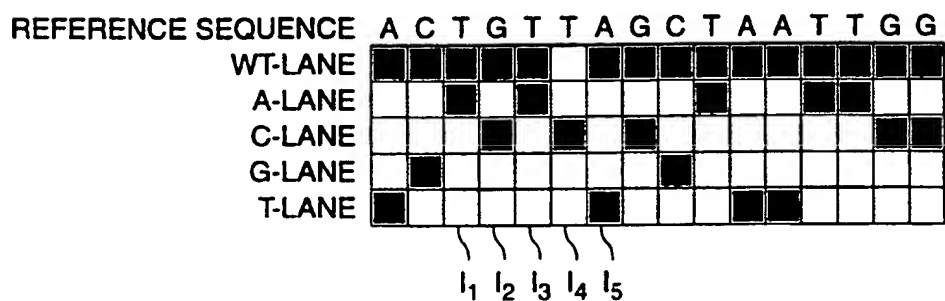


FIG. 8

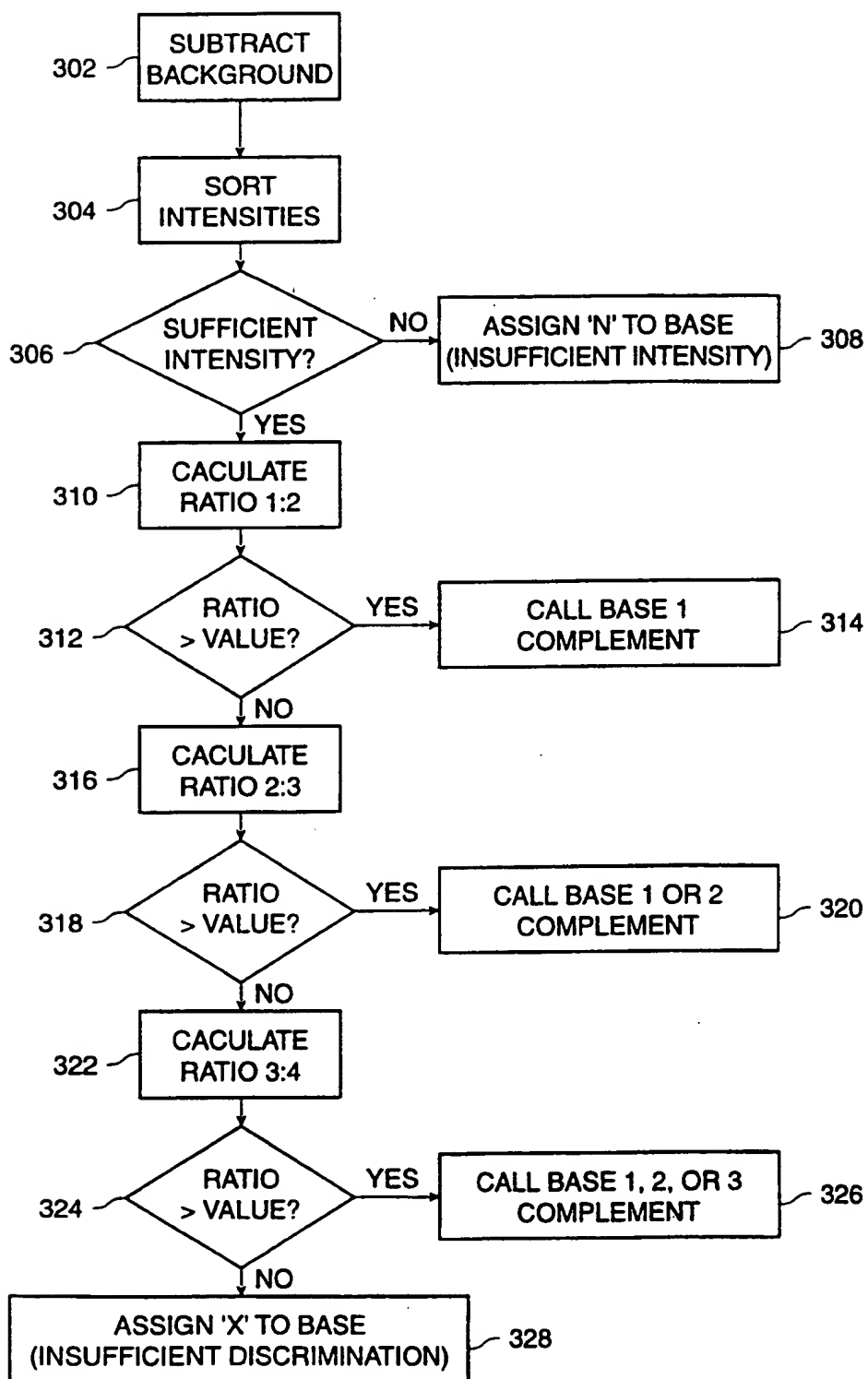


FIG. 9

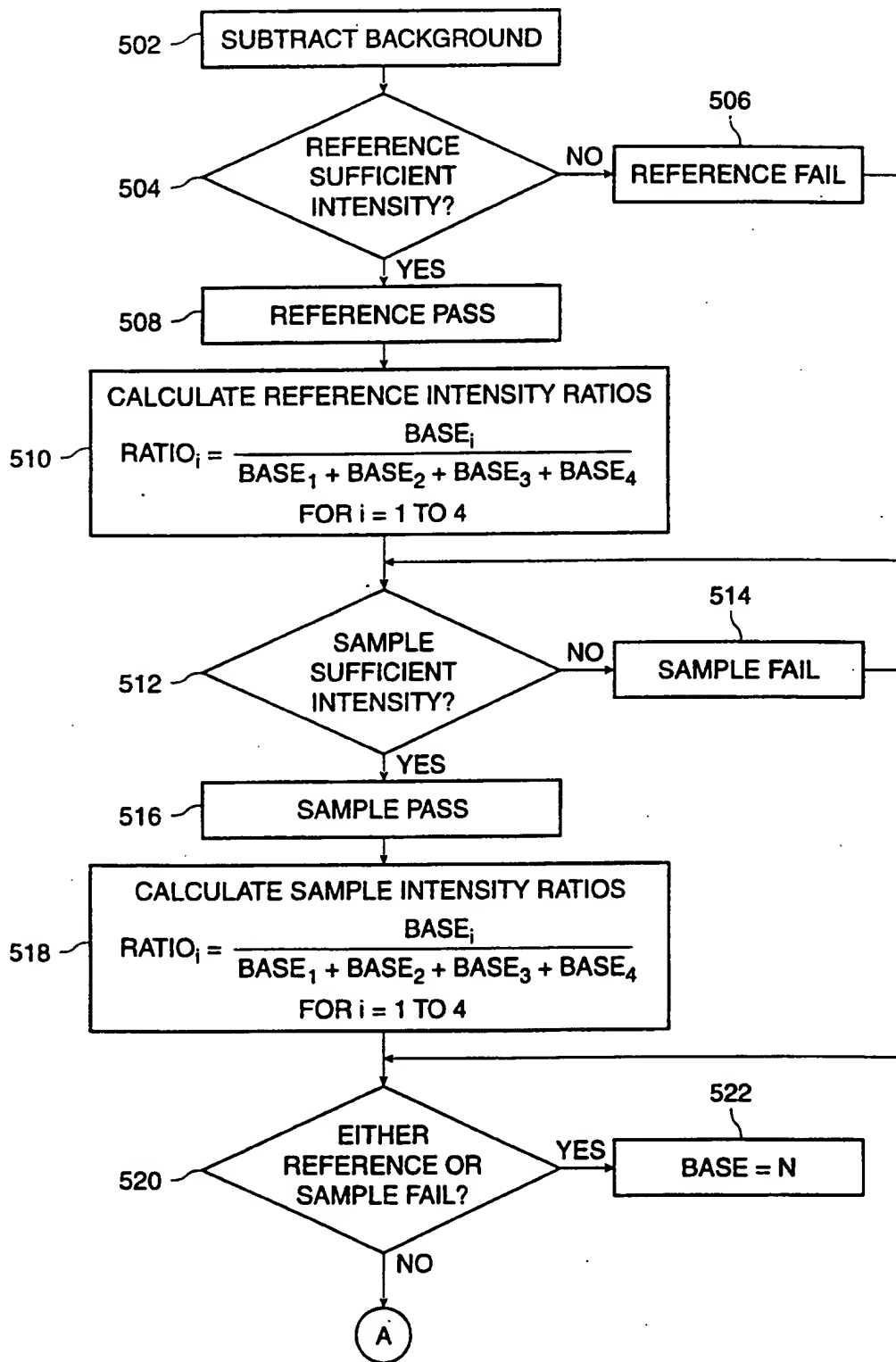


FIG. 10A

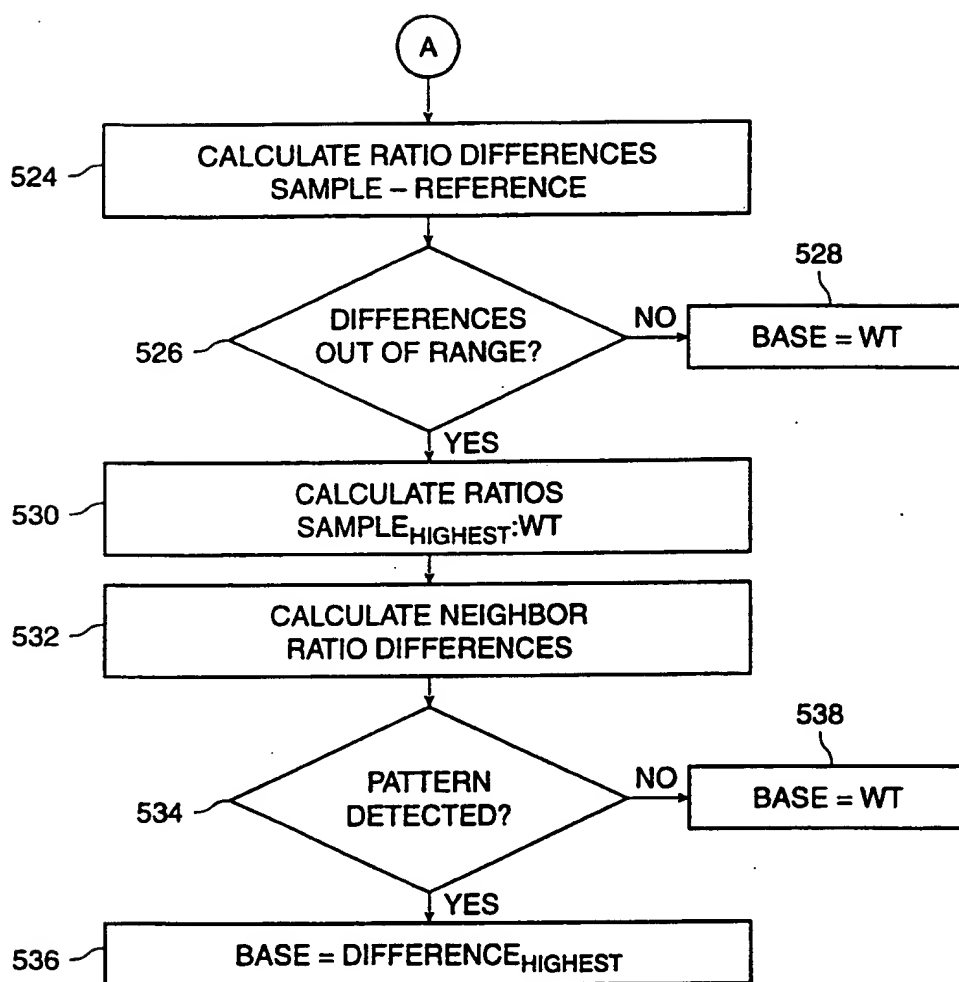


FIG. 10A
(CONTINUED)

POSITION	WT	REFERENCE				SAMPLE				RATIO OF RATIOS							
		BACK-GROUND	A	C	G	T	BACK-GROUND	A	C	G	T	A/A	C/C	G/G	T/T	BASE	CONFIDENCE
463	C	P	7.2	9.9	1.0	5.6	P	6.4	2.3	1.0	14.5	1.1	4.3	1.0	0.4	G	1
1		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

FIG. 10B

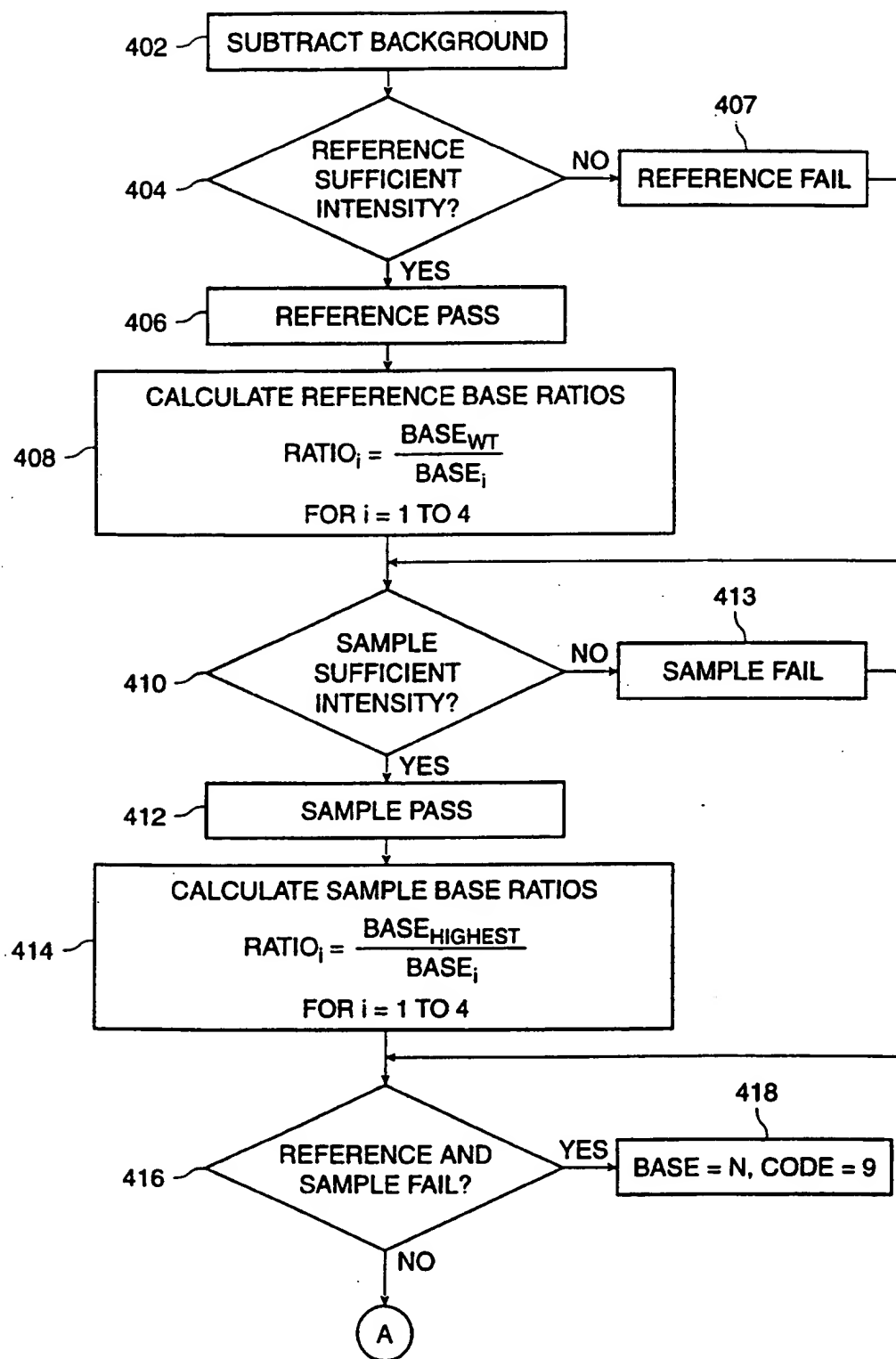


FIG. 11A

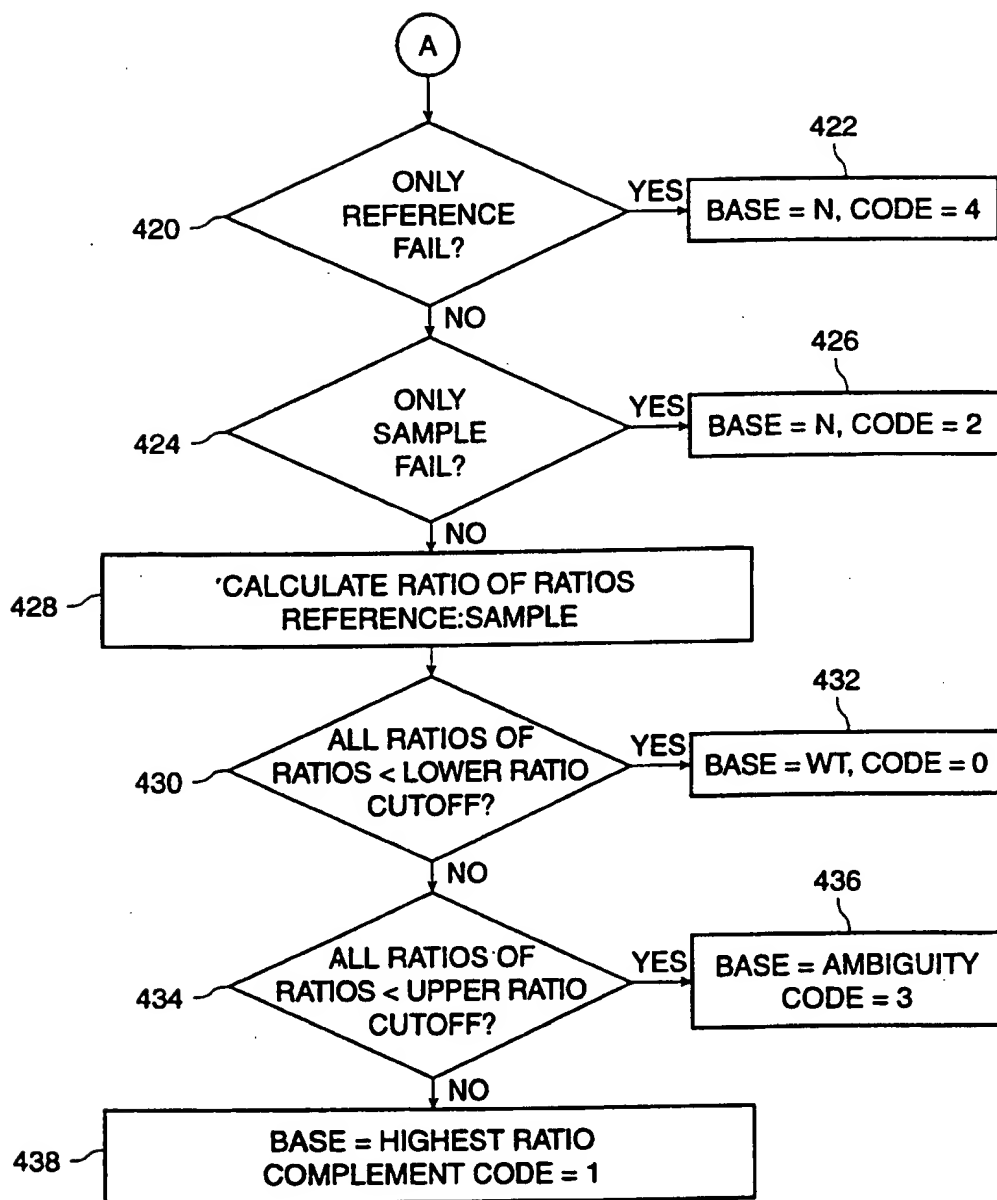


FIG. 11A
(CONTINUED)

BCK SUBTRACTED INTENSITIES		-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
RY090203.CQ1																		
POSITION:		234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250
WILDTYPE:		A	A	A	C	C	C	A	A	T	C	C	A	C	A	T	C	A
CALLED:		A	A	A	C	C	C	A	A	T	C	C	A	C	A	T	C	M
A		148	193	165	17	70	38	282	385	97	31	18	158	15	223	178	126	154
C		57	100	42	167	345	278	38	99	139	249	249	13	244	28	257	250	175
G		26	32	20	16	64	17	27	107	100	13	9	11	10	30	142	59	55
T		9	15	10	6	41	14	27	79	261	6	2	1	7	16	320	52	37
S		240	340	238	207	522	347	374	671	598	298	279	182	276	298	896	487	421
WTR		148	193	165	167	345	278	282	385	261	249	249	158	244	223	320	250	154
MAXR		148	193	165	167	345	278	282	385	261	249	249	158	244	223	320	250	175
MC090407.CQ1																		
POSITION:																		
WILDTYPE:		234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250
CALLED:		A	A	A	C	C	C	A	A	T	C	C	A	C	A	T	C	A
A		M	M	A	C	X	C	A	X	X	C	C	A	C	A	X	X	M
C		194	238	150	44	191	126	283	332	234	58	49	242	25	337	286	180	256
G		209	291	74	202	337	277	74	199	175	259	288	27	376	65	379	324	234
T		92	72	34	29	114	52	65	571	231	30	17	16	47	71	254	104	109
S		25	39	16	11	96	29	68	205	267	11	8	5	23	57	427	97	85
WTR		520	639	274	286	738	484	489	1307	906	357	362	291	472	529	1346	705	684
MAXR		194	238	150	202	337	277	283	332	267	259	288	242	376	337	427	324	256
		209	291	150	202	337	277	283	571	267	259	288	242	376	337	427	324	256

FIG. 11B

WTE/WTR	1.31	1.23	0.91	1.21	0.98	1.00	1.00	0.86	1.02	1.04	1.15	1.54	1.54	1.51	1.34	1.30	1.66
MAXE/WTR	1.42	1.51	0.91	1.21	0.98	1.00	1.00	1.48	1.02	1.04	1.15	1.54	1.54	1.51	1.34	1.30	1.66
N-L + N-R		0.79	-0.63	0.54	-0.25	0.01	0.14	0.94	0.14	-0.10	-0.27	0.38	0.04	0.14	-0.13	-0.40	
N-L		0.09	-0.60	0.30	-0.24	0.02	0.01	0.48	-0.46	0.02	0.11	0.38	0.01	-0.04	-0.17	-0.04	
N-R		0.60	-0.30	0.24	-0.02	-0.01	-0.48	0.46	-0.02	-0.11	-0.38	-0.01	0.04	0.17	0.04	-0.36	
N-L D(N-R)			-0.90	0.54	-0.25	0.01	-0.48	0.94	-0.48	-0.10	-0.27	0.38	0.04	0.14	-0.13		
N-R D(N-L)			-0.90	0.54	-0.25	0.01	-0.48	0.94	-0.48	-0.10	-0.27	0.38	0.04	0.14	-0.13		
L(N-L) - (N-R)L			0.29	0.07	0.22	0.02	0.49	0.02	0.44	0.13	0.50	0.39	0.03	0.21	0.21		
A+B-C			-2.10	1.01	-0.73	0.00	-1.44	1.86	-1.40	-0.33	-1.03	0.36	0.06	0.06	-0.48		
SUM MT/ SUM WT																	
INTENSITIES																	
N/L + N/R	2.16	1.88	1.15	1.39	1.41	1.39	1.31	1.95	1.52	1.20	1.30	1.60	1.71	1.78	1.50	1.45	1.63
N-L + N-R		2.50	1.45	2.18	2.04	2.05	1.61	2.77	2.04	1.71	1.89	2.18	2.03	2.22	1.88	1.85	
N-L		0.22	-0.48	0.10	0.02	0.03	-0.36	0.54	-0.06	-0.21	-0.10	0.10	0.02	0.17	-0.11	-0.12	
N-R		-0.28	-0.73	0.21	0.03	-0.02	-0.09	0.54	-0.43	-0.32	0.10	0.30	0.10	0.07	-0.27	-0.06	
		0.73	-0.23	-0.03	0.02	0.09	-0.64	0.43	0.32	-0.10	-0.30	-0.10	-0.07	0.27	0.06	-0.18	

FIG. 11B
(CONTINUED)

FIG. 11B
(CONTINUED)

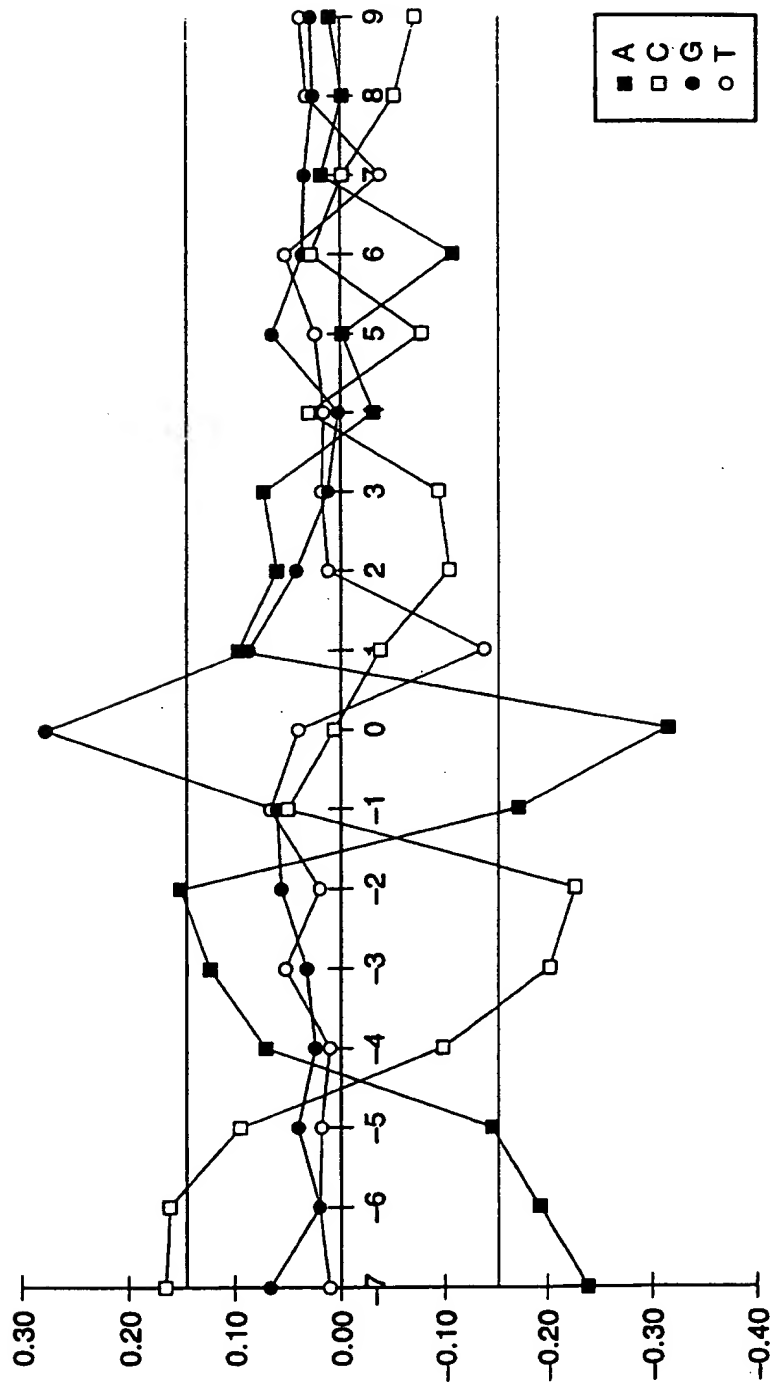


FIG. 11C

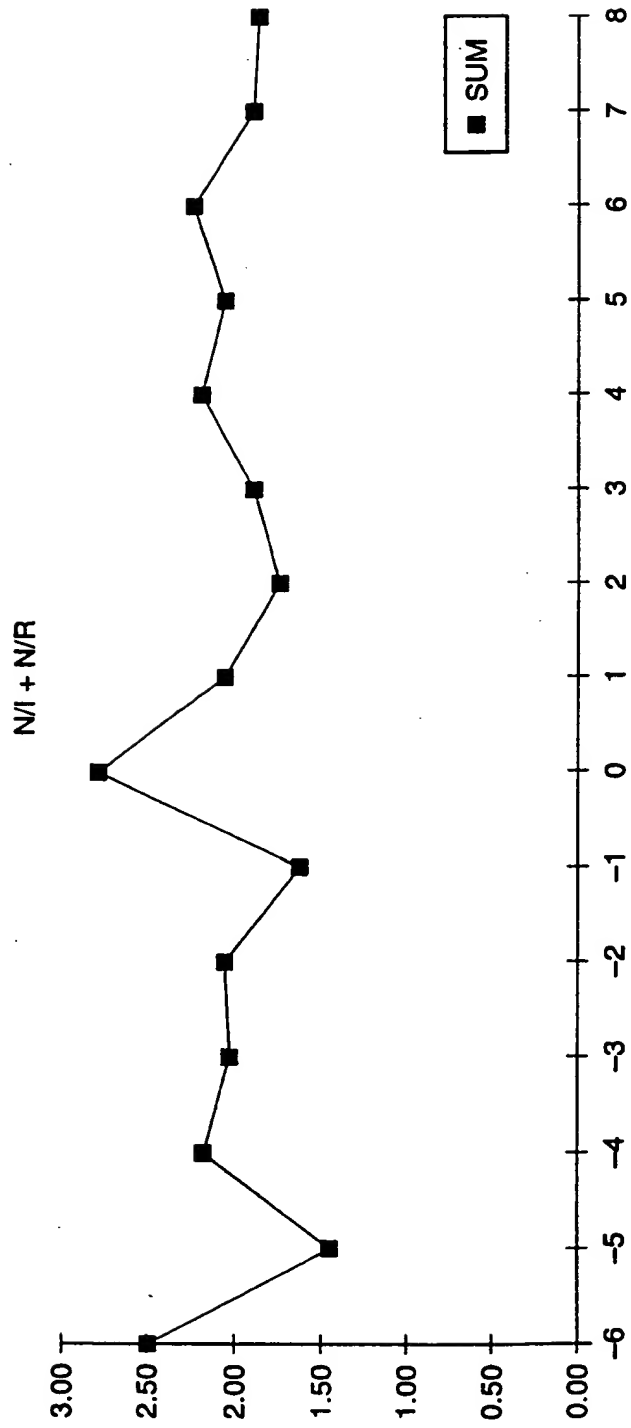


FIG. 11D.

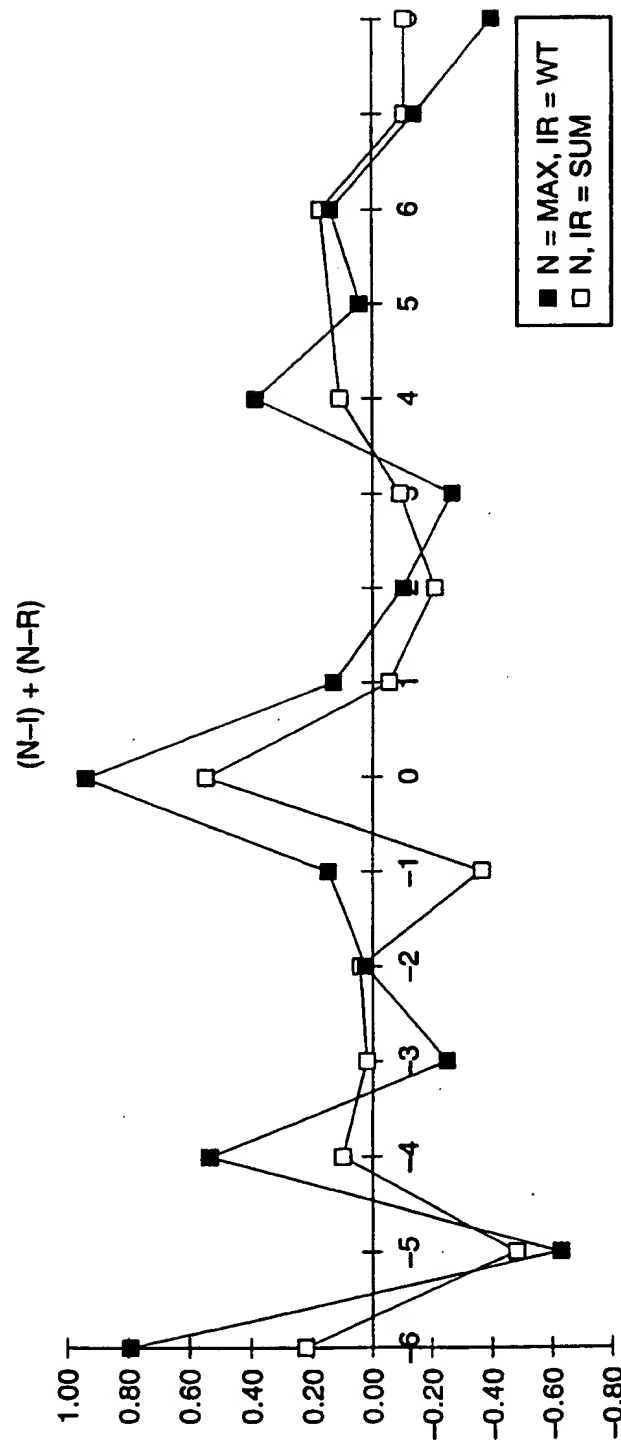


FIG. 11D
(CONTINUED)

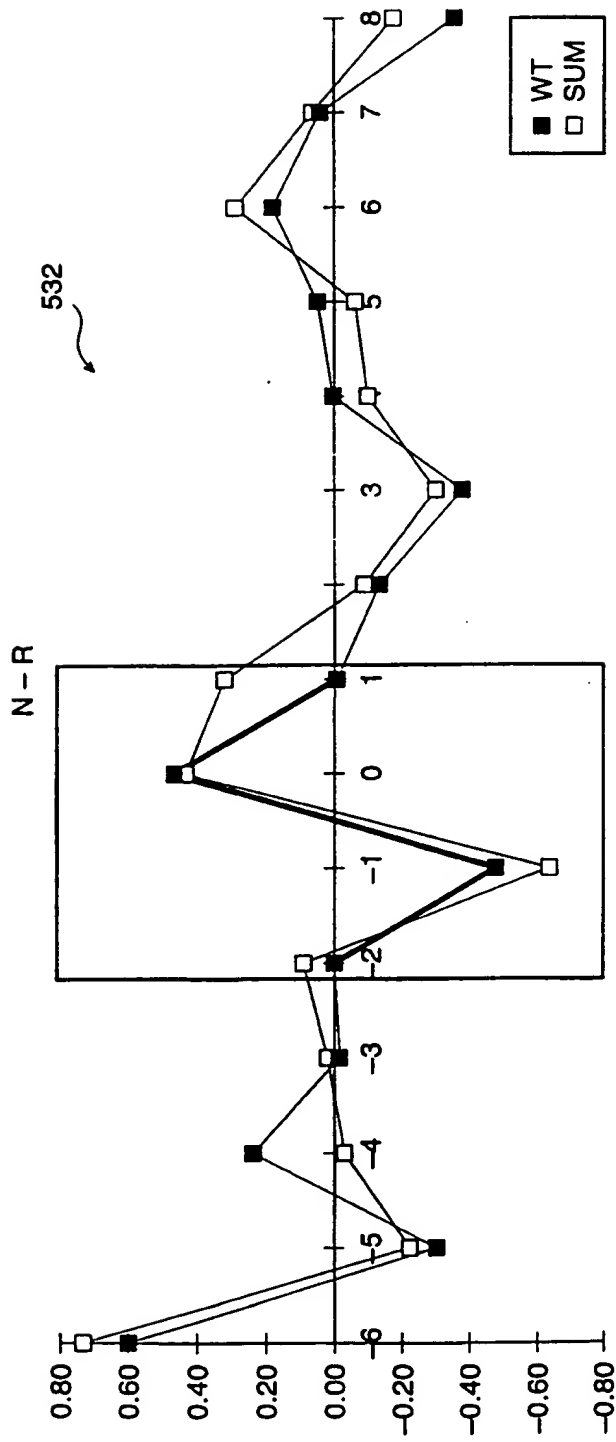


FIG. 11D
(CONTINUED)

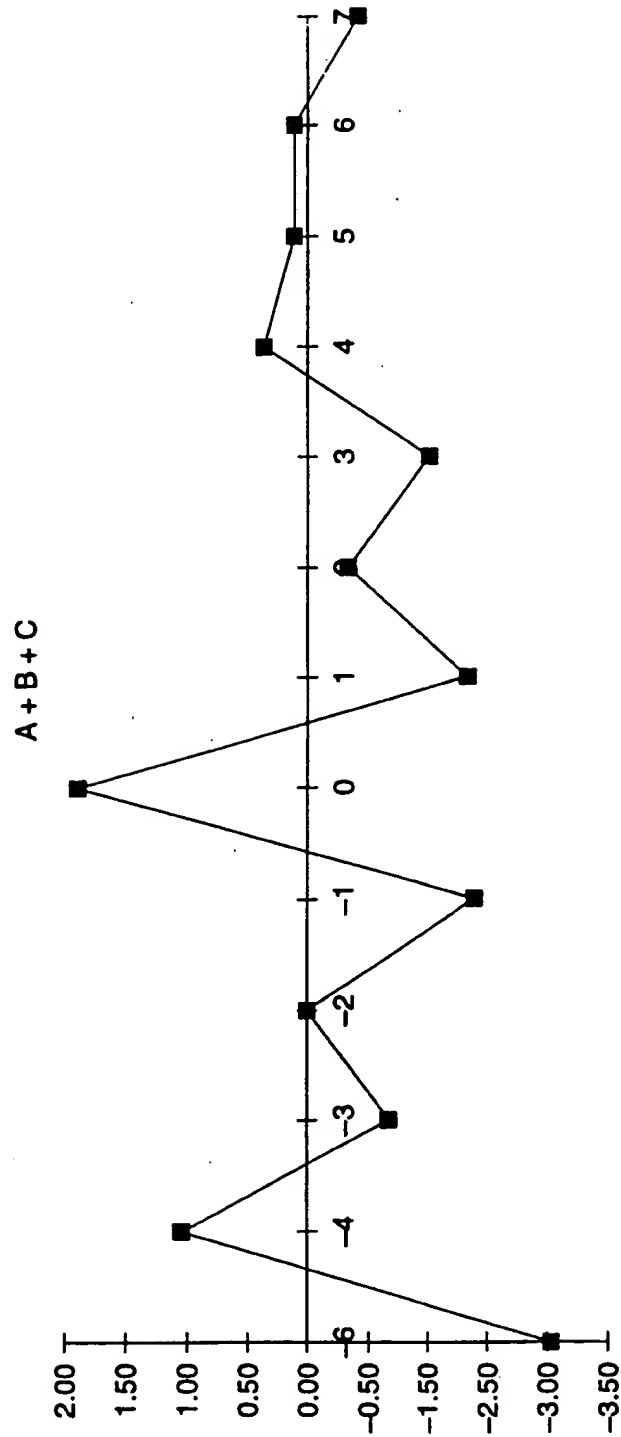


FIG. 11D
(CONTINUED)

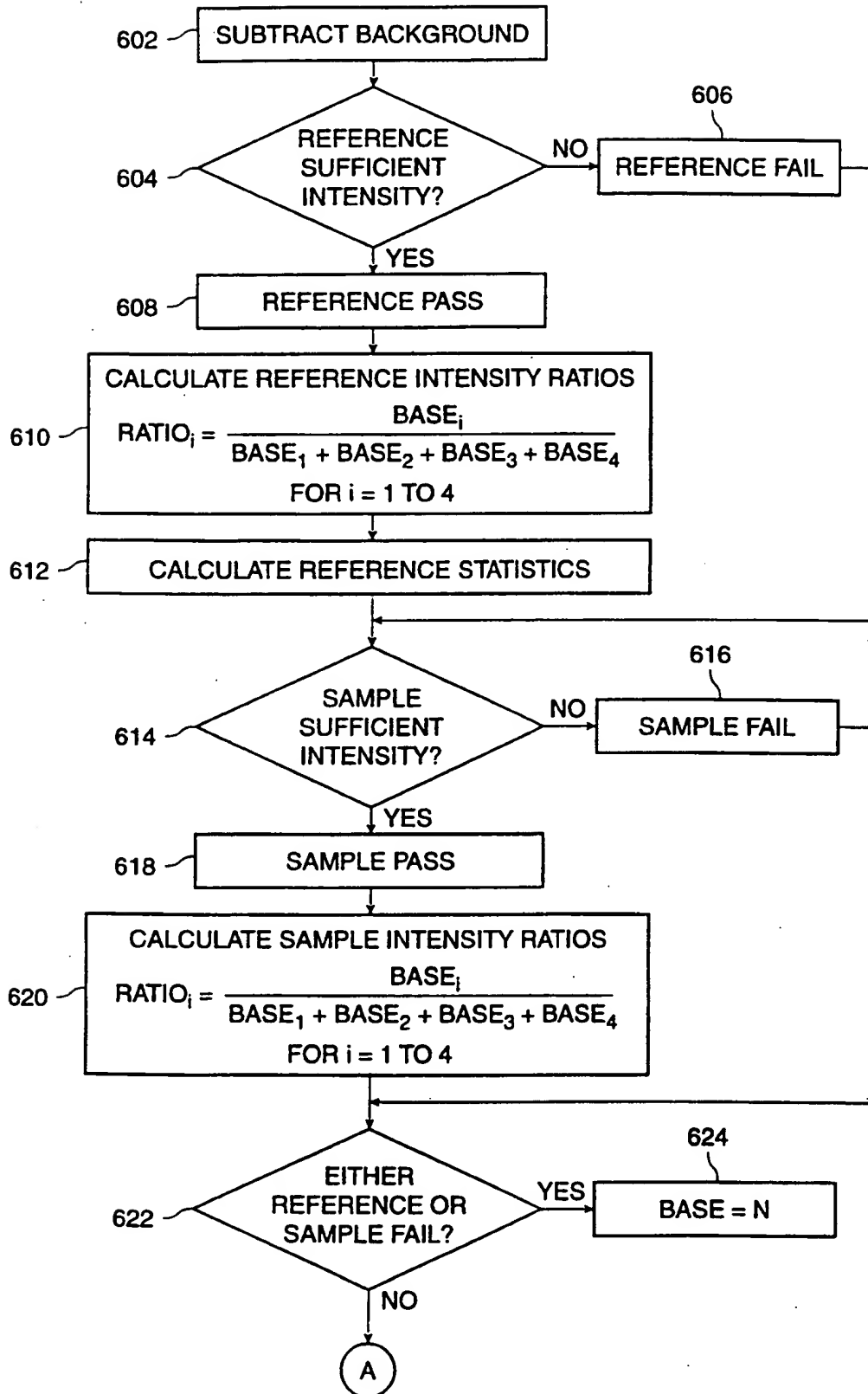


FIG. 12

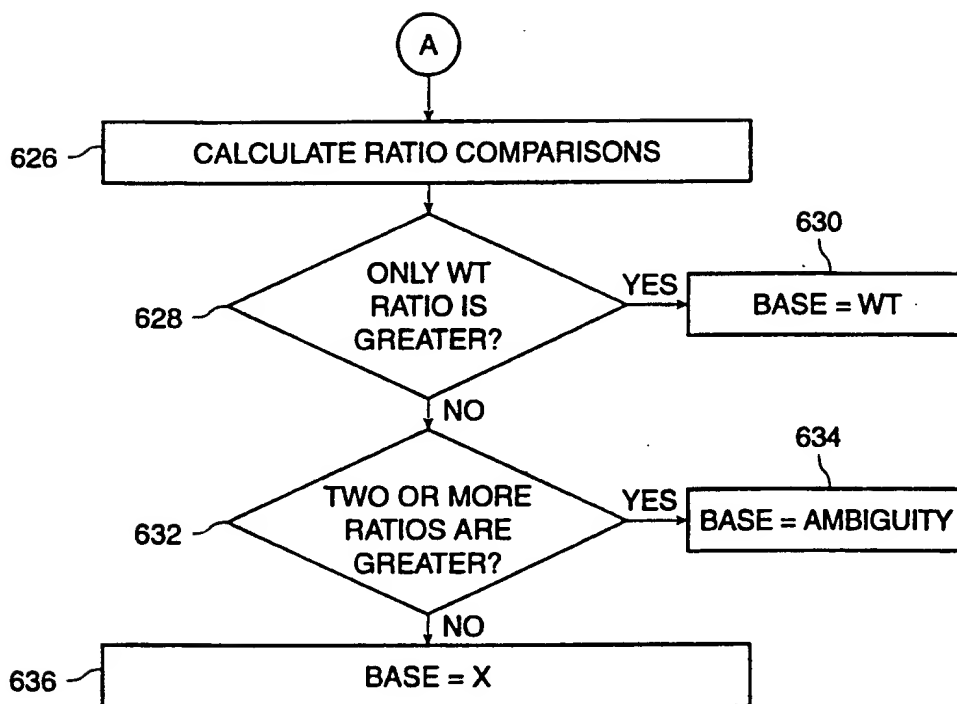


FIG. 12
(CONTINUED)

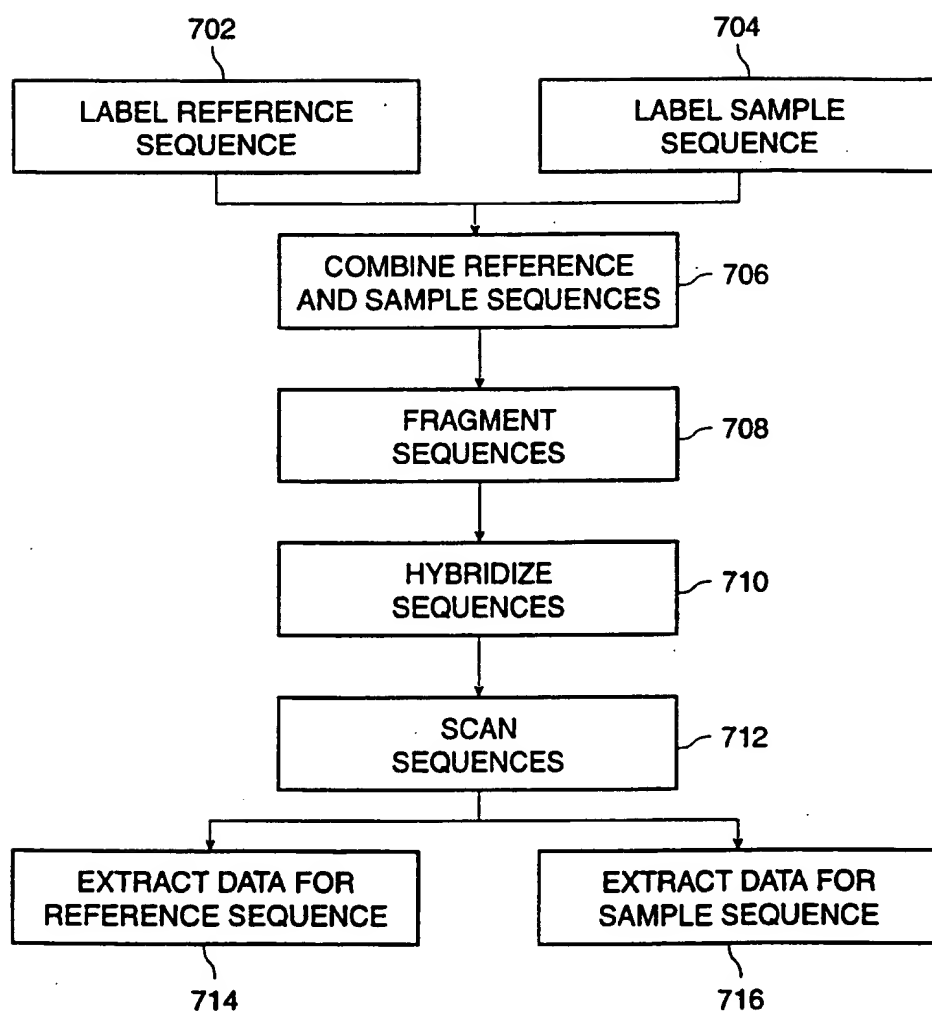


FIG. 13

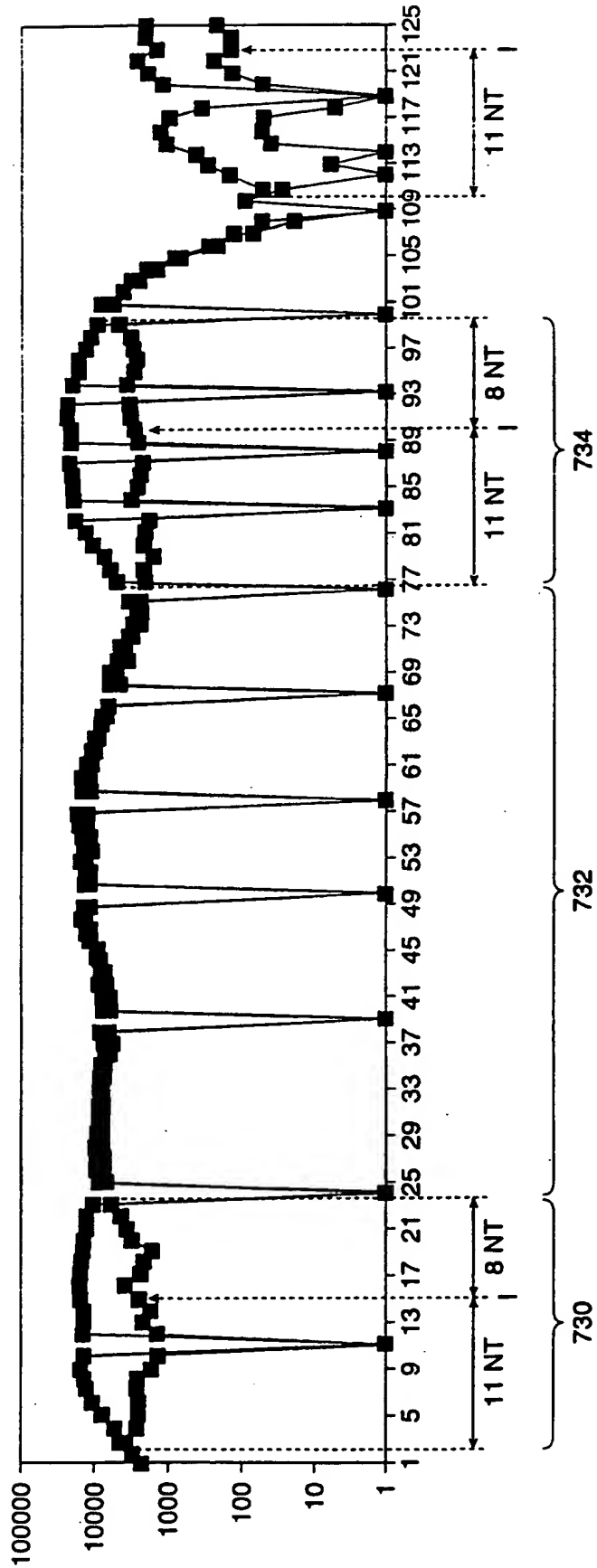


FIG. 14A

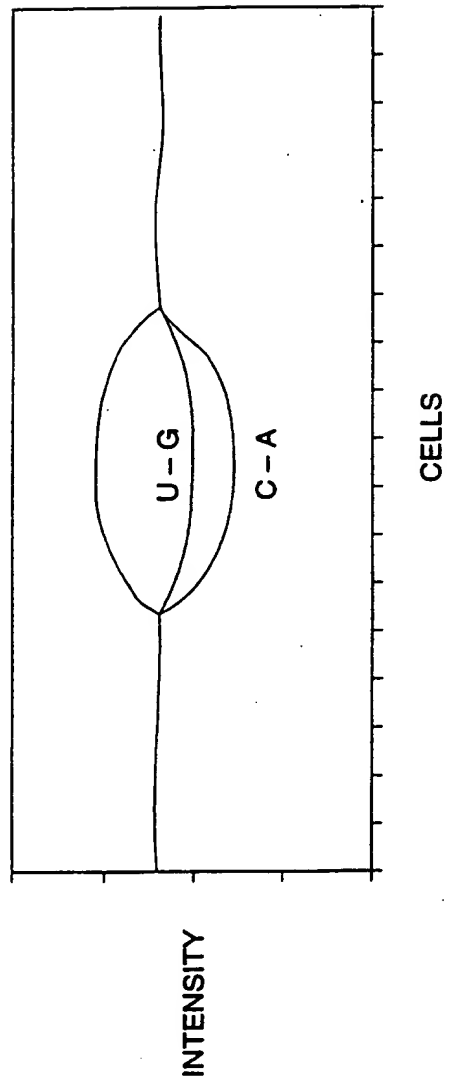


FIG. 14B

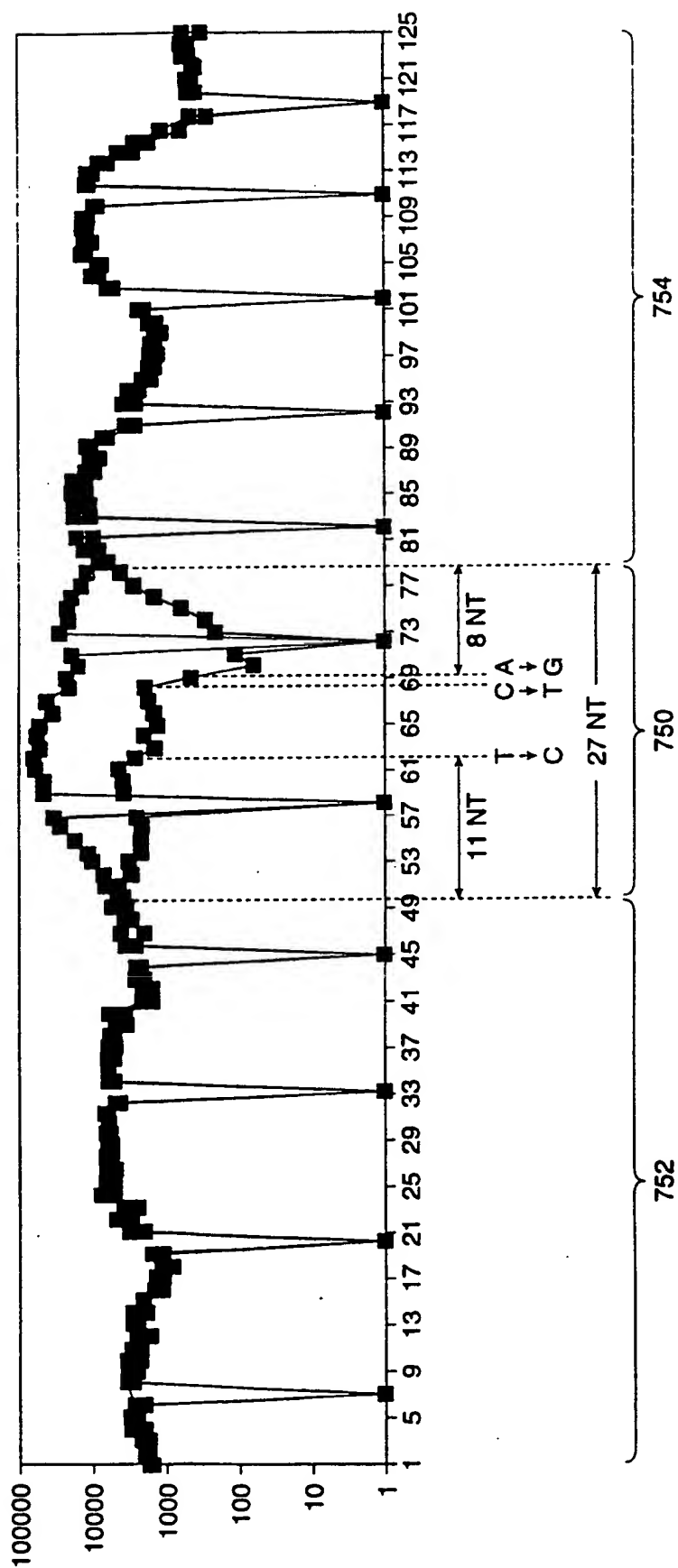


FIG. 14C

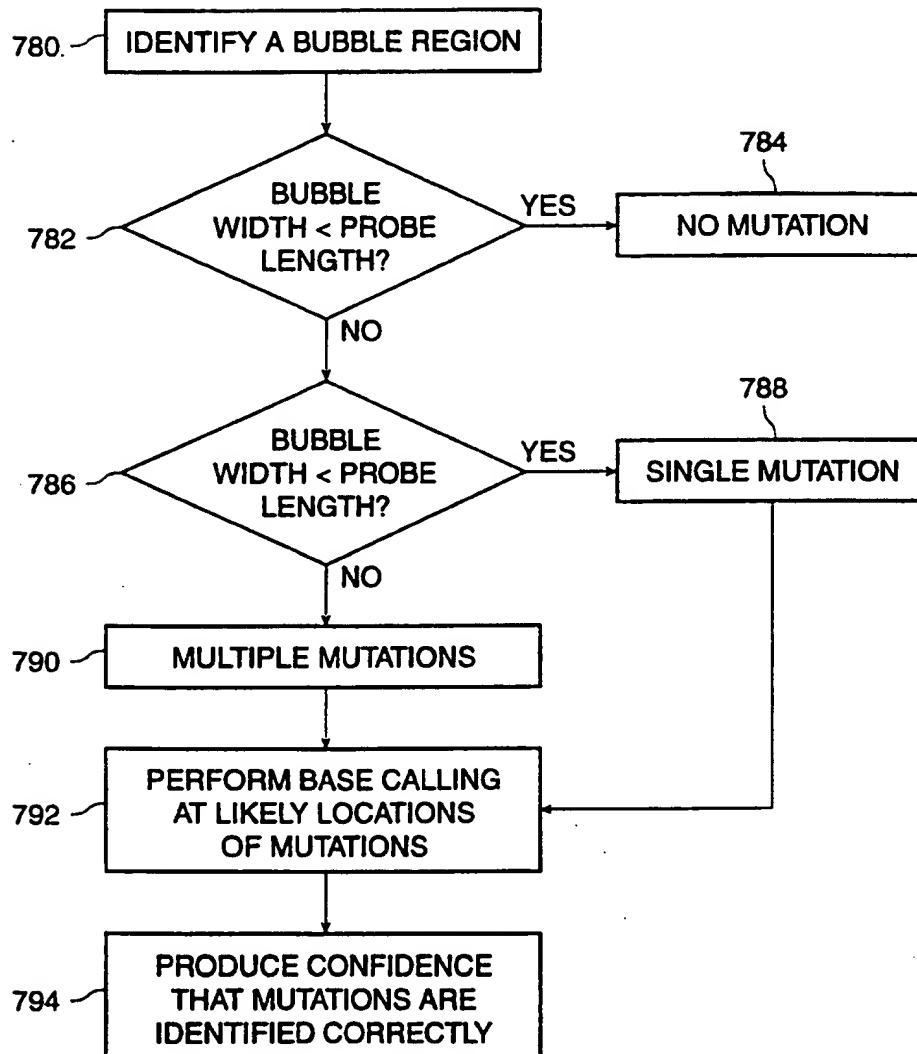


FIG. 15

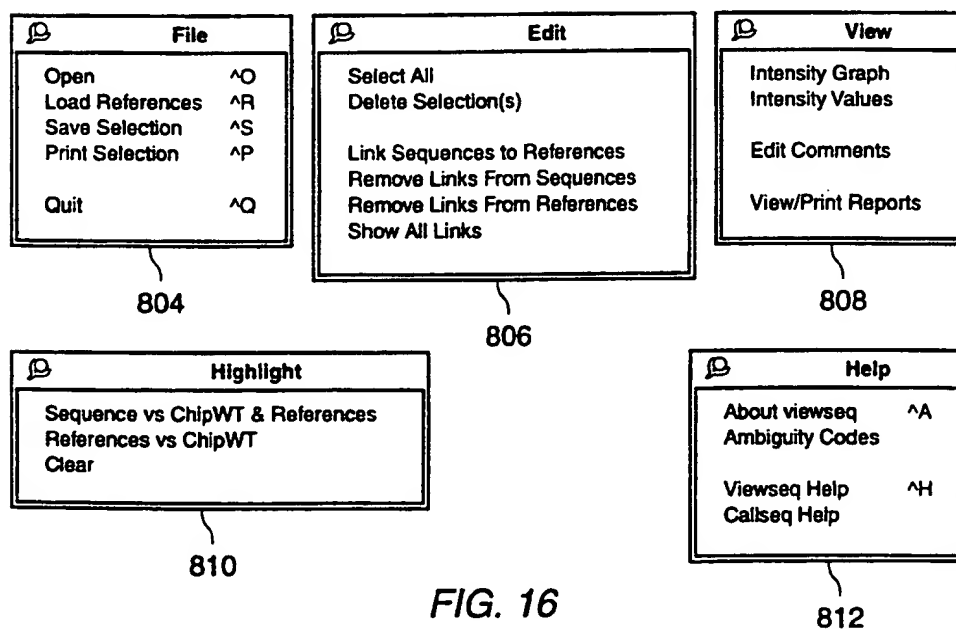
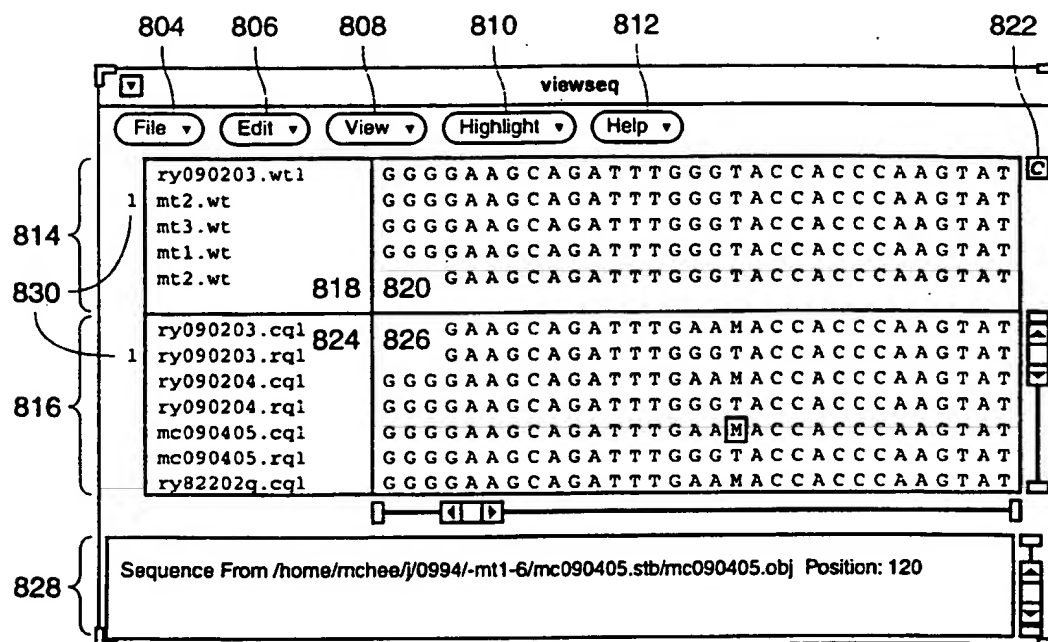


FIG. 16

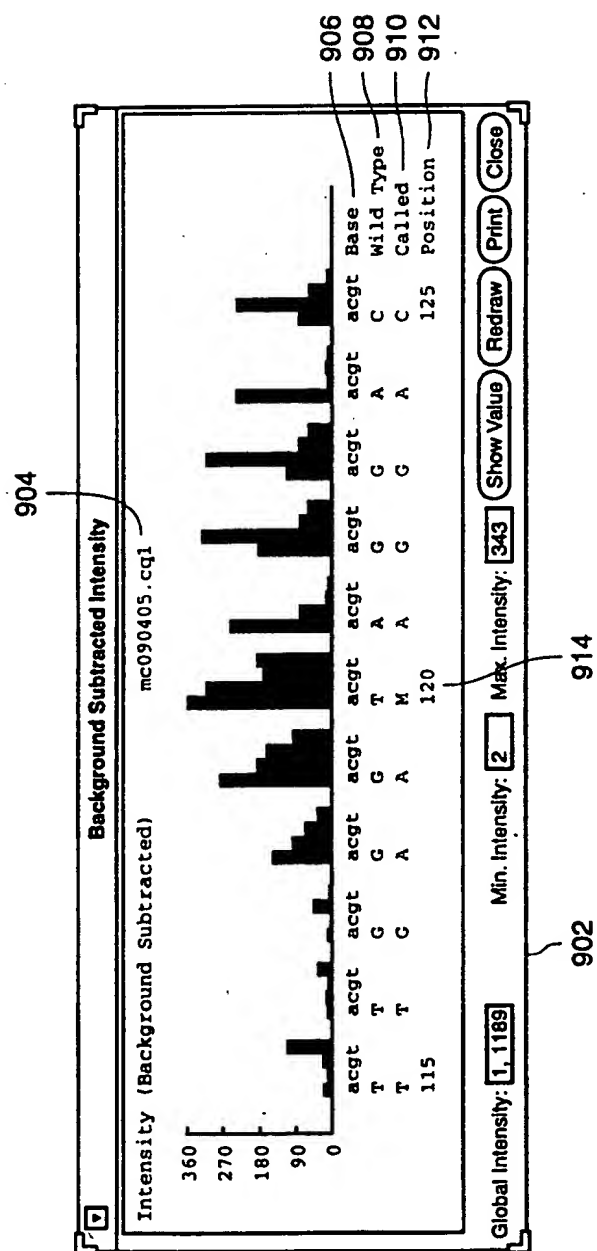


FIG. 17

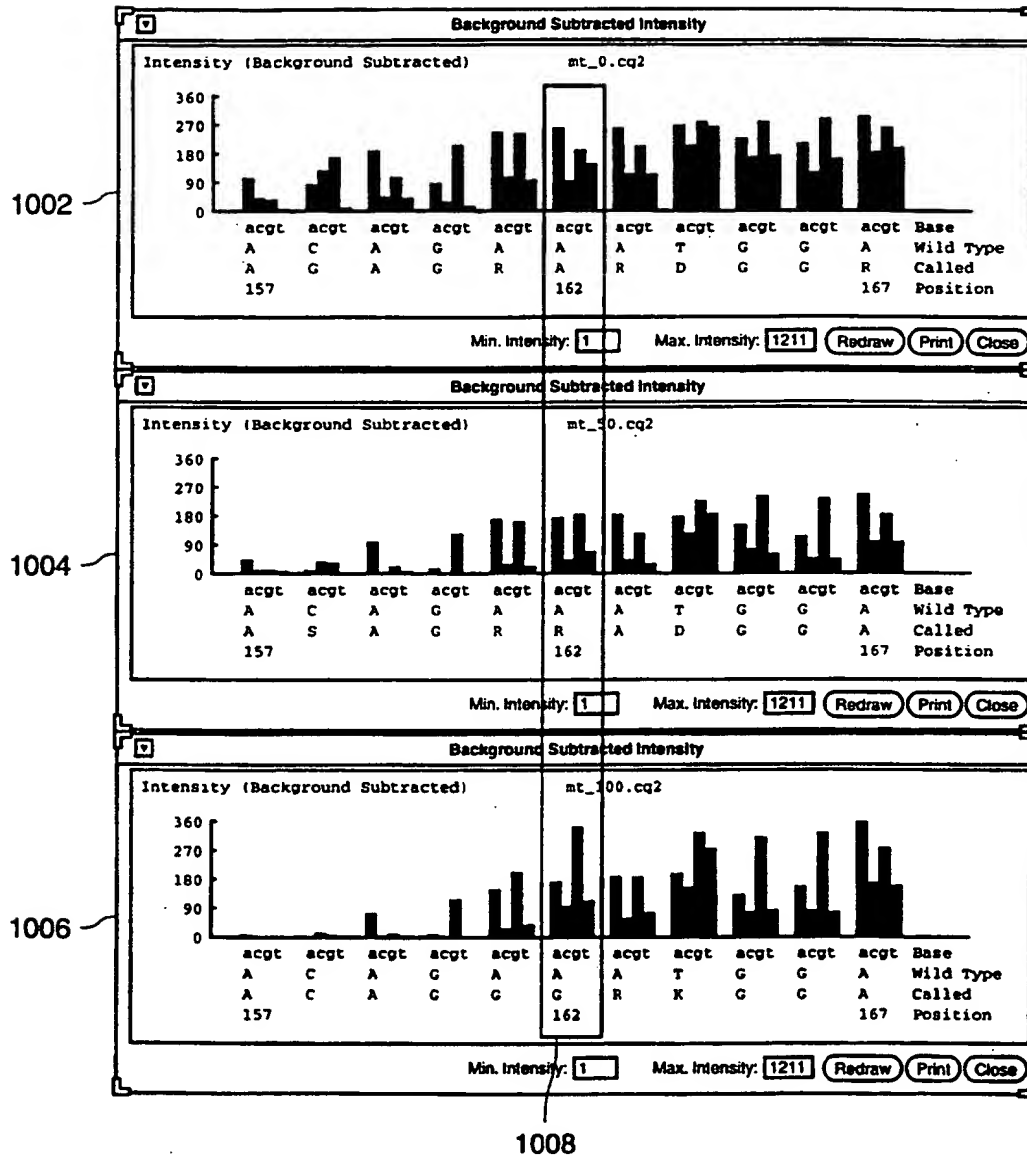


FIG. 18

	viewsseq		
File ▾	Edit ▾	View ▾	Help ▾
mt_0.wt2	GCATTAGTAGAGATATGTACAGAA	ATGGAAAAGGAAGGAAAAATTTCAAAAAATTGGGCC	C
1602.wt2	gcattagtagaaatttgtacaga	atggaaaagggaagggaatttcaaaaaattgggccc	
mt_0.cq2	GCATTAGTAGAGATATGTACAGAA	RDGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	
mt_10.cq2	GCATTAGTAGAGATATGKASAGRA	RDGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	
mt_25.cq2	GCATTAGTAGAGATATGKASAGRA	RDGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	
mt_50.cq2	GCATTAGTAGAGATATGTASAGRA	ADGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	
mt_75.cq2	GCATTAGTAGAGATATGTASAGRA	AGGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	
mt_90.cq2	GCATTAGTAGAGATATGTASAGRA	AGGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	
mt_100.cq2	GCATTAGTAGAGGNNNGACAGG	GKGGRAAXXXAAGGGAATAATNNNAAAAATTGGGCC	

1104
1106
1108
1110
1112
1114
1116
1118
1120

1108

FIG. 19

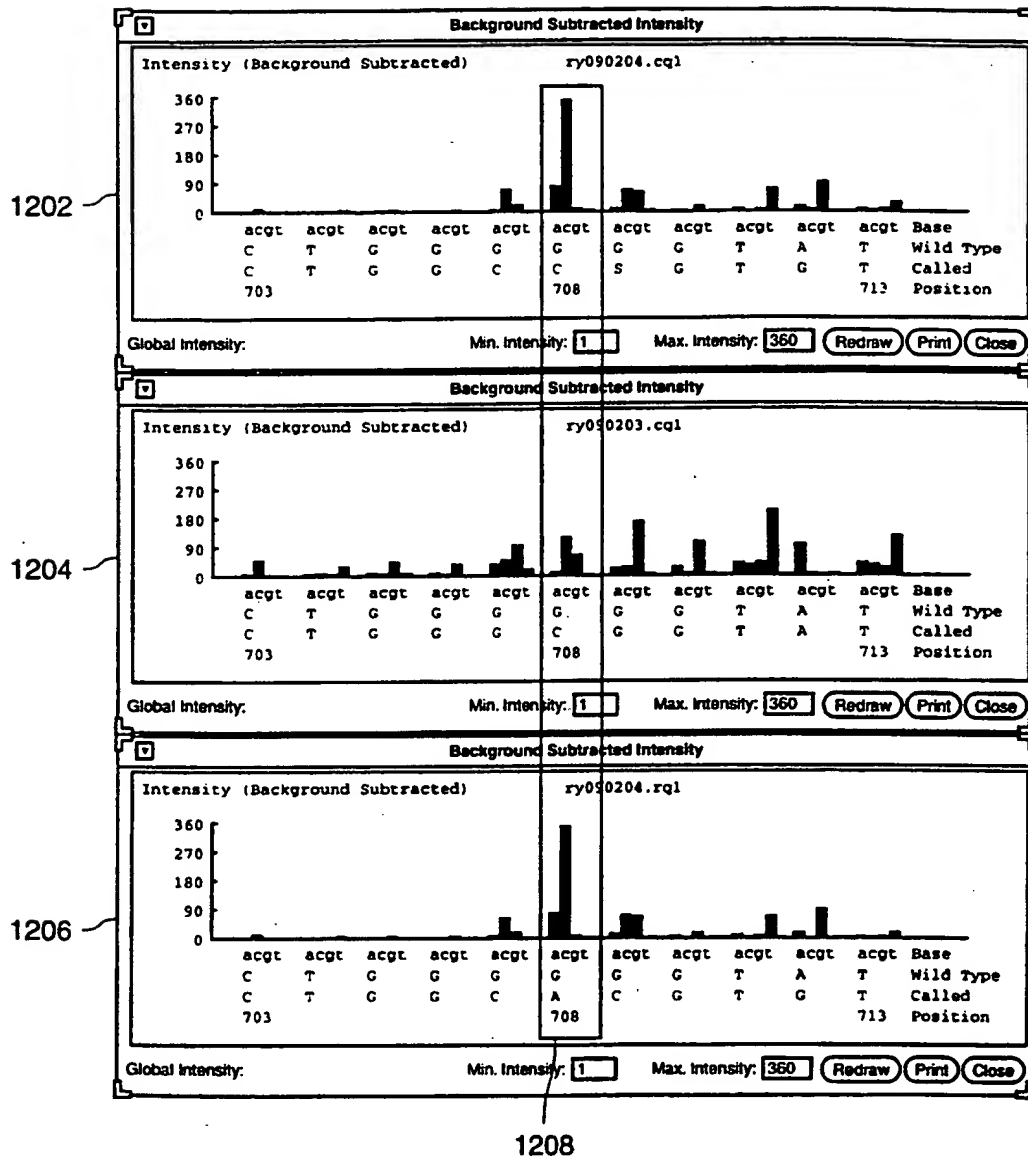


FIG. 20



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C12Q 1/68, C07H 21/02, 21/04		A1	(11) International Publication Number: WO 95/11995
			(43) International Publication Date: 4 May 1995 (04.05.95)
(21) International Application Number: PCT/US94/12305		Street, Mountain View, CA 94043 (US). HUBBELL, Earl, A. [US/US]; 1929 Crisanto #425, Mountain View, CA 94040 (US). LIPSHUTZ, Robert, J. [US/US]; 970 Palo Alto Avenue, Palo Alto, CA 94301 (US). LOBBAN, Peter, E. [US/US]; 273 Lowell Avenue, Palo Alto, CA 94301 (US). MIYADA, Charles, Garrett [US/US]; Sunnyvale, CA (US). MORRIS, MacDonald, S. [US/US]; P.O. Box 720488, San Jose, CA 95172 (US). SHAH, Nila [IN/US]; 12135 Saraglen, Saratoga, CA 95070 (US). SHELDON, Edward, L. [US/US]; 2031 Ashton Avenue, Menlo Park, CA 94025 (US). (74) Agents: LIEBESCHUETZ, Joseph et al.; Townsend and Townsend Khourie and Crew, Steuart Street Tower, 20th floor, One Market Plaza, San Francisco, CA 94105 (US). (81) Designated States: AM, AT, AU, BB, BG, BR, BY, CA, CH, CN, CZ, DE, DK, EE, ES, FI, GB, GE, HU, JP, KE, KG, KP, KR, KZ, LK, LR, LT, LU, LV, MD, MG, MN, MW, NL, NO, NZ, PL, PT, RO, RU, SD, SE, SI, SK, TJ, TT, UA, US, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ). Published <i>With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.</i>	
(22) International Filing Date: 26 October 1994 (26.10.94)			
(30) Priority Data: 08/143,312 26 October 1993 (26.10.93) US 08/284,064 2 August 1994 (02.08.94) US			
(60) Parent Application or Grant (63) Related by Continuation US Filed on 08/284,064 (CIP) 2 August 1994 (02.08.94)			
(71) Applicant (for all designated States except US): AFFYMAX TECHNOLOGIES N.V. [NL/NL]; De Ruyderkade 62, Curacao (AN).			
(72) Inventors; and (75) Inventors/Applicants (for US only): CHEE, Mark [US/US]; 3199 Waverly Street, Palo Alto, CA 94306 (US). CRONIN, Maureen, T. [US/US]; 771 Anderson Drive, Los Altos, CA 94024 (US). FODOR, Stephen, P., A. [US/US]; 3863 Nathan Way, Palo Alto, CA 94303 (US). GINGERAS, Thomas, R. [US/US]; 1568 Vista Club Circle, Santa Clara, CA 95054 (US). HUANG, Xiaohua, C. [-/US]; 937 Jackson			
(54) Title: ARRAYS OF NUCLEIC ACID PROBES ON BIOLOGICAL CHIPS			
(57) Abstract			
The invention provides chips of immobilized probes, and methods employing the chips, for comparing a reference polynucleotide sequence of known sequence with a target sequence showing substantial similarity with the reference sequence, but differing in the presence of e.g., mutations.			

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	GB	United Kingdom	MR	Mauritania
AU	Australia	GE	Georgia	MW	Malawi
BB	Barbados	GN	Guinea	NE	Niger
BE	Belgium	GR	Greece	NL	Netherlands
BF	Burkina Faso	HU	Hungary	NO	Norway
BG	Bulgaria	IE	Ireland	NZ	New Zealand
BJ	Benin	IT	Italy	PL	Poland
BR	Brazil	JP	Japan	PT	Portugal
BY	Belarus	KE	Kenya	RO	Romania
CA	Canada	KG	Kyrgyzstan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LU	Luxembourg	TD	Chad
CS	Czechoslovakia	LV	Latvia	TG	Togo
CZ	Czech Republic	MC	Monaco	TJ	Tajikistan
DE	Germany	MD	Republic of Moldova	TT	Trinidad and Tobago
DK	Denmark	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	US	United States of America
FI	Finland	MN	Mongolia	UZ	Uzbekistan
FR	France			VN	Viet Nam
GA	Gabon				

ARRAYS OF NUCLEIC ACID PROBES ON BIOLOGICAL CHIPS5 Cross-Reference to Related Application

This application is a continuation-in-part of USSN 08/284,064, filed August 2, 1994, which is a continuation-in-part of USSN 08/143,312, filed October 26, 1993, each of which is incorporated by reference in its entirety for all purposes. Research leading to the invention was funded in part by NIH grant No. 1R01HG00813-01, and the government may have certain rights to the invention.

Background of the Invention15 Field of the Invention

The present invention provides arrays of oligonucleotide probes immobilized in microfabricated patterns on silica chips for analyzing molecular interactions of biological interest. The invention therefore relates to diverse fields impacted by the nature of molecular interaction, including chemistry, biology, medicine, and medical diagnostics.

Description of Related Art

Oligonucleotide probes have long been used to detect complementary nucleic acid sequences in a nucleic acid of interest (the "target" nucleic acid). In some assay formats, the oligonucleotide probe is tethered, i.e., by covalent attachment, to a solid support, and arrays of oligonucleotide probes immobilized on solid supports have been used to detect specific nucleic acid sequences in a target nucleic acid. See, e.g., PCT patent publication Nos. WO 89/10977 and 89/11548. Others have proposed the use of large numbers of oligonucleotide probes to provide the complete nucleic acid sequence of a target nucleic acid but failed to provide an enabling method for using arrays of immobilized probes for this purpose. See U.S. Patent Nos. 5,202,231 and 5,002,867 and PCT patent publication No. WO 93/17126.

The development of VLSIPS™ technology has provided methods for making very large arrays of oligonucleotide probes in very small arrays. See U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, each of which is incorporated herein by reference. U.S. Patent application Serial No. 082,937, filed June 25, 1993, describes methods for making arrays of oligonucleotide probes that can be used to provide the complete sequence of a target nucleic acid and to detect the presence of a nucleic acid containing a specific nucleotide sequence.

Microfabricated arrays of large numbers of oligonucleotide probes, called "DNA chips" offer great promise for a wide variety of applications. New methods and reagents are required to realize this promise, and the present invention helps meet that need.

SUMMARY OF THE INVENTION

The invention provides several strategies employing immobilized arrays of probes for comparing a reference sequence of known sequence with a target sequence showing substantial similarity with the reference sequence, but differing in the presence of, e.g., mutations. In a first embodiment, the invention provides a tiling strategy employing an array of immobilized oligonucleotide probes comprising at least two sets of probes. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. A second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. The probes in the first probe set have at

least two interrogation positions corresponding to two contiguous nucleotides in the reference sequence. One interrogation position corresponds to one of the contiguous nucleotides, and the other interrogation position to the other.

In a second embodiment, the invention provides a tiling strategy employing an array comprising four probe sets. A first probe set comprises a plurality of probes, each probe comprising a segment of at least three nucleotides exactly complementary to a subsequence of the reference sequence, the segment including at least one interrogation position complementary to a corresponding nucleotide in the reference sequence. Second, third and fourth probe sets each comprise a corresponding probe for each probe in the first probe set. The probes in the second, third and fourth probe sets are identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the at least one interrogation position, except that the at least one interrogation position is occupied by a different nucleotide in each of the four corresponding probes from the four probe sets. The first probe set often has at least 100 interrogation positions corresponding to 100 contiguous nucleotides in the reference sequence. Sometimes the first probe set has an interrogation position corresponding to every nucleotide in the reference sequence. The segment of complementarity within the probe set is usually about 9-21 nucleotides. Although probes may contain leading or trailing sequences in addition to the 9-21 sequences, many probes consist exclusively of a 9-21 segment of complementarity.

In a third embodiment, the invention provides immobilized arrays of probes tiled for multiple reference sequences. One such array comprises at least one pair of first and second probe groups, each group comprising first and second sets of probes as defined in the first embodiment. Each probe in the first probe set from the first group is exactly complementary to a subsequence of a first reference sequence, and each probe in the first probe set from the second group is exactly

complementary to a subsequence of a second reference sequence. Thus, the first group of probes are tiled with respect to a first reference sequence and the second group of probes with respect to a second reference sequence. Each group of probes
5 can also include third and fourth sets of probes as defined in the second embodiment. In some arrays of this type, the second reference sequence is a mutated form of the first reference sequence.

In a fourth embodiment, the invention provides arrays for
10 block tiling. Block tiling is a species of the general tiling strategies described above. The usual unit of a block tiling array is a group of probes comprising a wildtype probe, a first set of three mutant probes and a second set of three mutant probes. The wildtype probe comprises a segment of at
15 least three nucleotides exactly complementary to a subsequence of a reference sequence. The segment has at least first and second interrogation positions corresponding to first and second nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to a
20 sequence comprising the wildtype probe or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the first interrogation position, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes
25 in the second set of three mutant probes are each identical to a sequence comprising the wildtype probes or a subsequence of at least three nucleotides thereof including the first and second interrogation positions, except in the second interrogation position, which is occupied by a different
30 nucleotide in each of the three mutant probes and the wildtype probe.

In a fifth embodiment, the invention provides methods of comparing a target sequence with a reference sequence using
arrays of immobilized pooled probes. The arrays employed in
35 these methods represent a further species of the general tiling arrays noted above. In these methods, variants of a reference sequence differing from the reference sequence in at least one nucleotide are identified and each is assigned a

designation. An array of pooled probes is provided, with each pool occupying a separate cell of the array. Each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular designation.

5 The array is then contacted with a target sequence comprising a variant of the reference sequence. The relative hybridization intensities of the pools in the array to the target sequence are determined. The identity of the target sequence is deduced from the pattern of hybridization intensities. Often, each variant is assigned a designation having at least one digit and at least one value for the digit. In this case, each pool comprises a probe comprising a segment exactly complementary to each variant sequence assigned a particular value in a particular digit. When 15 variants are assigned successive numbers in a numbering system of base m having n digits, $n \times (m-1)$ pooled probes are used are used to assign each variant a designation.

In a sixth embodiment, the invention provides a pooled probe for trellis tiling, a further species of the general 20 tiling strategy. In trellis tiling, the identity of a nucleotide in a target sequence is determined from a comparison of hybridization intensities of three pooled trellis probes. A pooled trellis probe comprises a segment exactly complementary to a subsequence of a reference sequence except at a first interrogation position occupied by a pooled 25 nucleotide N , a second interrogation position occupied by a pooled nucleotide selected from the group of three consisting of (1) M or K , (2) R or Y and (3) S or W , and a third interrogation position occupied by a second pooled nucleotide selected from the group. The pooled nucleotide occupying the 30 second interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the second pooled probe and reference sequence are maximally aligned, and the pooled nucleotide occupying the 35 third interrogation position comprises a nucleotide complementary to a corresponding nucleotide from the reference sequence when the third pooled probe and the reference

sequence are maximally aligned. Standard IUPAC nomenclature is used for describing pooled nucleotides.

In trellis tiling, an array comprises at least first, second and third cells, respectively occupied by first, second and third pooled probes, each according to the generic description above. However, the segment of complementarity, location of interrogation positions, and selection of pooled nucleotide at each interrogation position may or may not differ between the three pooled probes subject to the following constraint. One of the three interrogation positions in each of the three pooled probes must align with the same corresponding nucleotide in the reference sequence. This interrogation position must be occupied by a N in one of the pooled probes, and a different pooled nucleotide in each of the other two pooled probes.

In a seventh embodiment, the invention provides arrays for bridge tiling. Bridge tiling is a species of the general tiling strategies noted above, in which probes from the first probe set contain more than one segment of complementarity. In bridge tiling, a nucleotide in a reference sequence is usually determined from a comparison of four probes. A first probe comprises at least first and second segments, each of at least three nucleotides and each exactly complementary to first and second subsequences of a reference sequences. The segments including at least one interrogation position corresponding to a nucleotide in the reference sequence. Either (1) the first and second subsequences are noncontiguous in the reference sequence, or (2) the first and second subsequences are contiguous and the first and second segments are inverted relative to the first and second subsequences. The arrays further comprises second, third and fourth probes, which are identical to a sequence comprising the first probe or a subsequence thereof comprising at least three nucleotides from each of the first and second segments, except in the at least one interrogation position, which differs in each of the probes. In a species of bridge tiling, referred to as deletion tiling, the first and second subsequences are separated by one or two nucleotides in the reference sequence.

In an eighth embodiment, the invention provides arrays of probes for multiplex tiling. Multiplex tiling is a strategy, in which the identity of two nucleotides in a target sequence is determined from a comparison of the hybridization intensities of four probes, each having two interrogation positions. Each of the probes comprising a segment of at least 7 nucleotides that is exactly complementary to a subsequence from a reference sequence, except that the segment may or may not be exactly complementary at two interrogation positions. The nucleotides occupying the interrogation positions are selected by the following rules: (1) the first interrogation position is occupied by a different nucleotide in each of the four probes, (2) the second interrogation position is occupied by a different nucleotide in each of the four probes, (3) in first and second probes, the segment is exactly complementary to the subsequence, except at no more than one of the interrogation positions, (4) in third and fourth probes, the segment is exactly complementary to the subsequence, except at both of the interrogation positions.

In a ninth embodiment, the invention provides arrays of immobilized probes including helper mutations. Helper mutations are useful for, e.g., preventing self-annealing of probes having inverted repeats. In this strategy, the identity of a nucleotide in a target sequence is usually determined from a comparison of four probes. A first probe comprises a segment of at least 7 nucleotides exactly complementary to a subsequence of a reference sequence except at one or two positions, the segment including an interrogation position not at the one or two positions. The one or two positions are occupied by helper mutations. Second, third and fourth mutant probes are each identical to a sequence comprising the wildtype probe or a subsequence thereof including the interrogation position and the one or two positions, except in the interrogation position, which is occupied by a different nucleotide in each of the four probes.

In a tenth embodiment, the invention provides arrays of probes comprising at least two probe sets, but lacking a probe set comprising probes that are perfectly matched to a

reference sequence. Such arrays are usually employed in methods in which both reference and target sequence are hybridized to the array. The first probe set comprising a plurality of probes, each probe comprising a segment exactly complementary to a subsequence of at least 3 nucleotides of a reference sequence except at an interrogation position. The second probe set comprises a corresponding probe for each probe in the first probe set, the corresponding probe in the second probe set being identical to a sequence comprising the corresponding probe from the first probe set or a subsequence of at least three nucleotides thereof that includes the interrogation position, except that the interrogation position is occupied by a different nucleotide in each of the two corresponding probes and the complement to the reference sequence.

In an eleventh embodiment, the invention provides methods of comparing a target sequence with a reference sequence comprising a predetermined sequence of nucleotides using any of the arrays described above. The methods comprise hybridizing the target nucleic acid to an array and determining which probes, relative to one another, in the array bind specifically to the target nucleic acid. The relative specific binding of the probes indicates whether the target sequence is the same or different from the reference sequence. In some such methods, the target sequence has a substituted nucleotide relative to the reference sequence in at least one undetermined position, and the relative specific binding of the probes indicates the location of the position and the nucleotide occupying the position in the target sequence. In some methods, a second target nucleic acid is also hybridized to the array. The relative specific binding of the probes then indicates both whether the target sequence is the same or different from the reference sequence, and whether the second target sequence is the same or different from the reference sequence. In some methods, when the array comprises two groups of probes tiled for first and second reference sequences, respectively, the relative specific binding of probes in the first group indicates whether the

target sequence is the same or different from the first reference sequence. The relative specific binding of probes in the second group indicates whether the target sequence is the same or different from the second reference sequence.

5 Such methods are particularly useful for analyzing heterologous alleles of a gene. Some methods entail hybridizing both a reference sequence and a target sequence to any of the arrays of probes described above. Comparison of the relative specific binding of the probes to the reference
10 and target sequences indicates whether the target sequence is the same or different from the reference sequence.

In a twelfth embodiment, the invention provides arrays of immobilized probes in which the probes are designed to tile a reference sequence from a human immunodeficiency virus.

15 Reference sequences from either the reverse transcriptase gene or protease gene of HIV are of particular interest. Some chips further comprise arrays of probes tiling a reference sequence from a 16S RNA or DNA encoding the 16S RNA from a pathogenic microorganism. The invention further provides
20 methods of using such arrays in analyzing a HIV target sequence. The methods are particularly useful where the target sequence has a substituted nucleotide relative to the reference sequence in at least one position, the substitution conferring resistance to a drug use in treating a patient
25 infected with a HIV virus. The methods reveal the existence of the substituted nucleotide. The methods are also particularly useful for analyzing a mixture of undetermined proportions of first and second target sequences from different HIV variants. The relative specific binding of
30 probes indicates the proportions of the first and second target sequences.

In a thirteenth embodiment, the invention provides arrays of probes tiled based on reference sequence from a CFTR gene. A preferred array comprises at least a group of probes
35 comprising a wildtype probe, and five sets of three mutant probes. The wildtype probe is exactly complementary to a subsequence of a reference sequence from a cystic fibrosis gene, the segment having at least five interrogation positions

corresponding to five contiguous nucleotides in the reference sequence. The probes in the first set of three mutant probes are each identical to the wildtype probe, except in a first of the five interrogation positions, which is occupied by a
5 different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the second set of three mutant probes are each identical to the wildtype probe, except in a second of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant
10 probes and the wildtype probe. The probes in the third set of three mutant probes are each identical to the wildtype probe, except in a third of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the
15 fourth set of three mutant probes are each identical to the wildtype probe, except in a fourth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. The probes in the fifth set of three mutant probes are each identical to
20 the wildtype probe, except in a fifth of the five interrogation positions, which is occupied by a different nucleotide in each of the three mutant probes and the wildtype probe. Preferably, a chip comprises two such groups of probes. The first group comprises a wildtype probe exactly
25 complementary to a first reference sequence, and the second group comprises a wildtype probe exactly complementary to a second reference sequence that is a mutated form of the first reference sequence.

The invention further provides methods of using the
30 arrays of the invention for analyzing target sequences from a CFTR gene. The methods are capable of simultaneously analyzing first and second target sequences representing heterozygous alleles of a CFTR gene.

In a fourteenth embodiment, the invention provides arrays
35 of probes tiling a reference sequence from a p53 gene, an hMLH1 gene and/or an MSH2 gene. The invention further provides methods of using the arrays described above to

analyze these genes. The method are useful, e.g., for diagnosing patients susceptible to developing cancer.

In a fifteenth embodiment, the invention provides arrays of probes tiling a reference sequence from a mitochondrial genome. The reference sequence may comprise part or all of the D-loop region, or all, or substantially all, of the mitochondrial genome. The invention further provides method of using the arrays described above to analyze target sequences from a mitochondrial genome. The methods are useful for identifying mutations associated with disease, and for forensic, epidemiological and evolutionary studies.

BRIEF DESCRIPTION OF THE FIGURES

Fig. 1: Basic tiling strategy. The figure illustrates the relationship between an interrogation position (I) and a corresponding nucleotide (n) in the reference sequence, and between a probe from the first probe set and corresponding probes from second, third and fourth probe sets.

Fig. 2: Segment of complementarity in a probe from the first probe set.

Fig. 3: Incremental succession of probes in a basic tiling strategy. The figure shows four probe sets, each having three probes. Note that each probe differs from its predecessor in the same set by the acquisition of a 5' nucleotide and the loss of a 3' nucleotide, as well as in the nucleotide occupying the interrogation position.

Fig. 4: Exemplary arrangement of lanes on a chip. The chip shows four probe sets, each having five probes and each having a total of five interrogation positions (I1-I5), one per probe.

Fig. 5: Hybridization pattern of chip having probes laid down in lanes. Dark patches indicate hybridization. The probes in the lower part of the figure occur at the column of the array indicated by the arrow when the probes length is 15 and the interrogation position 7.

Fig. 6: Strategies for detecting deletion and insertion mutations. Bases in brackets may or may not be present.

Fig. 7: Block tiling strategy. The probe from the first probe set has three interrogation positions. The probes from the other probe sets have only one of these interrogation positions.

5 Fig. 8: Multiplex tiling strategy. Each probe has two interrogation positions.

Fig. 9. Helper mutation strategy. The segment of complementarity differs from the complement of the reference sequence at a helper mutation as well as the interrogation
10 position.

Fig. 10 Layout of probes on the HV 407 chip. The figure shows successive rows of sequence each of which is subdivided into four lanes. The four lanes correspond to the A-, C-, G- and T-lanes on the chip. Each probe is represented by the
15 nucleotide occupying its interrogation position. The letter "N" indicates a control probe or empty column. The different sized-probes are laid out in parallel. That is, from top-to-bottom, a row of 13 mers is followed by a row of 15 mers, which is followed by a row of 17 mers, which is followed by a
20 row of 19 mers.

Fig. 11 Fluorescence pattern of HV 407 hybridized to a target sequence (pPol19) identical to the chips reference sequence.

Fig. 12 Sequence read from HV 407 chip hybridized to
25 pPol19 and 4MUT18 (separate experiments). The reference sequence is designated "wildtype." Beneath the reference sequence are four rows of sequence read from the chip hybridized to the pPol19 target, the first row being read from 13 mers, the second row from 15 mers, the third row from 17
30 mers and the fourth row from 19 mers. Beneath these sequences, there are four further rows of sequence read from the chip hybridized to the HXB2 target. Successive rows are read from 13 mers, 15 mers, 17 mers and 19 mers. Each
35 nucleotide in a row is called from the relative fluorescence intensities of probes in A-, C-, G- and T-lanes. Regions of ambiguous sequence read from the chip are highlighted. The strain differences between the HXB2 sequence and the reference sequence that were correctly detected are indicated (*), and

those that could not be called are indicated (o). (The nucleotide at position 417 was read correctly in some experiments). The location of some mutations known to be associated with drug resistance that occur in readable regions of the chip are shown above (codon number) and below (mutant nucleotide) the sequence designated "wildtype." The locations of primer used to amplify the target sequence are indicated by arrows.

Fig. 13: Detection of mixed target sequences. The mutant target differs from the wildtype by a single mutation in codon 67 of the reverse transcriptase gene. Each different sized group of probes has a column of four probes for reading the nucleotide in which the mutation occurs. The four probes occupying a column are represented by a single probe in the figure with the symbol (o) indicating the interrogation position, which is occupied by a different nucleotide in each probe.

Fig. 14: Fluorescence intensities of target bound to 13 mers and 15 mers for different proportions of mutant and wildtype target. The fluorescence intensities are from probes having interrogation positions for reading the nucleotide at which the mutant and wildtype targets diverge.

Fig. 15: Sequence read from protease chip from four clinical samples before and after treatment with ddI>.

Fig. 16: Block tiling array of probes for analyzing a CFTR point mutation. Each probe show actually represents four probes, with one probe having each of A, C, G or T at the interrogation position N. In the order shown, the first probe shown on the left is tiled from the wildtype reference sequence, the second probe from the mutant sequence, and so on in alternating fashion. Note that all of the probes are identical except at the interrogation position, which shifts one position between successive probes tiled from the same reference sequence (e.g., the first, third and fifth probes in the left hand column.) The grid shows the hybridization intensities when the array is hybridized to the reference sequence.

Fig. 17: Hybridization pattern for heterozygous target. The figure shows the hybridization pattern when the array of the previous figure is hybridized to a mixture of mutant and wildtype reference sequences.

5 Fig. 18, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to a wild-type target; in panel C, the chip was hybridized to a mutant $\Delta F508$ target; and in panel B, the chip was hybridized to a mixture of the
10 wild-type and mutant targets.

 Fig. 19, in sheets 1 - 3, corresponding to panels A, B, and C of Fig. 18, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted
15 are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound
20 the target with the second highest intensity of all the substitution probes ("2nd Highest").

 Fig. 20, in panels A, B, and C, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to the wt480 target; in panel
25 C, the chip was hybridized to the mu480 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets.

 Fig. 21, in sheets 1 - 3, corresponding to panels A, B, and C of Fig. 20, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis
30 show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing
35 the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

Fig. 22, in panels A and B, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to nucleic acid derived from the genomic DNA of an individual with wild-type $\Delta F508$ sequences; in panel B, the target nucleic acid originated from a heterozygous (with respect to the $\Delta F508$ mutation) individual.

Fig. 23, in sheets 1 and 2, corresponding to panels A and B of Fig. 22, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

Fig. 24: Hybridization of homozygous wildtype (A) and heterozygous (B) target sequences from exon 11 of the CFTR gene to a block tiling array designed to detect G551D and Q552X mutations in CFTR gene.

Fig. 25: Hybridization of homozygous wildtype (A) and $\Delta F508$ mutant (B) target sequences from exon 10 of the CFTR gene to a block tiling array designed to detect mutations, $\Delta F508$, $\Delta I507$ and F508C.

Fig. 26: Hybridization of heterozygous mutant target sequences, $\Delta F508$ /F508C, to the array of Fig. 25.

Fig. 27 shows the alignment of some of the probes on a p53 DNA chip with a 12-mer model target nucleic acid.

Fig. 28 shows a set of 10-mer probes for a p53 exon 6 DNA chip.

Fig. 29 shows that very distinct patterns are observed after hybridization of p53 DNA chips with targets having different 1 base substitutions. In the first image in Fig. 29, the 12-mer probes that form perfect matches with the wild-type target are in the first row (top). The 12-mer probes with single base mismatches are located in the second, third, and fourth rows and have much lower signals.

Fig. 30, in graphs 2, 3, and 4, graphically depicts the data in Fig. 29. On each graph, the X ordinate is the position of the probe in its row on the chip, and the Y ordinate is the signal at that probe site after hybridization.

5 Fig. 31 shows the results of hybridizing mixed target populations of WT and mutant p53 genes to the p53 DNA chip.

Fig. 32, in graphs 1-4, shows (see Fig. 30 as well) the hybridization efficiency of a 10-mer probe array as compared to a 12-mer probe array.

10 Fig. 33 shows an image of a p53 DNA chip hybridized to a target DNA.

Fig. 34 illustrates how the actual sequence was read from the chip shown in Fig. 33. Gaps in the sequence of letters in the WT rows correspond to control probes or sites. Positions at which bases are miscalled are represented by letters in italic type in cells corresponding to probes in which the WT bases have been substituted by other bases.

15 Fig. 35 shows the human mitochondrial genome; "O_H" is the H strand origin of replication, and arrows indicate the cloned unshaded sequence.

20 Fig. 36 shows the image observed from application of a sample of mitochondrial DNA derived nucleic acid (from the mt4 sample) on a DNA chip.

Fig. 37 is similar to Fig. 36 but shows the image observed from the mt5 sample.

25 Fig. 38 shows the predicted difference image between the mt4 and mt5 samples on the DNA chip based on mismatches between the two samples and the reference sequence.

Fig. 39 shows the actual difference image observed for the mt4 and mt5 samples.

30 Fig. 40, in sheets 1 and 2, shows a plot of normalized intensities across rows 10 and 11 of the array and a tabulation of the mutations detected.

Fig. 41 shows the discrimination between wild-type and mutant hybrids obtained with the chip. A median of the six normalized hybridization scores for each probe was taken; the graph plots the ratio of the median score to the normalized

35

hybridization score versus mean counts. A ratio of 1.6 and mean counts above 50 yield no false positives.

Fig. 42 illustrates how the identity of the base mismatch may influence the ability to discriminate mutant and wild-type sequences more than the position of the mismatch within an oligonucleotide probe. The mismatch position is expressed as % of probe length from the 3'-end. The base change is indicated on the graph.

Fig. 43 provides a 5' to 3' sequence listing of one target corresponding to the probes on the chip. X is a control probe. Positions that differ in the target (i.e., are mismatched with the probe at the designated site) are in bold.

Fig. 44 shows the fluorescence image produced by scanning the chip described in Fig. 17 when hybridized to a sample.

Fig. 45 illustrates the detection of 4 transitions in the target sequence relative to the wild-type probes on the chip in Fig. 44.

Fig. 46: VLSIPS™ technology applied to the light directed synthesis of oligonucleotides. Light (hv) is shone through a mask (M_1) to activate functional groups (-OH) on a surface by removal of a protecting group (X). Nucleoside building blocks protected with photoremovable protecting groups (T-X, C-X) are coupled to the activated areas. By repeating the irradiation and coupling steps, very complex arrays of oligonucleotides can be prepared.

Fig. 47: Use of the VLSIPS™ process to prepare "nucleoside combinatorials" or oligonucleotides synthesized by coupling all four nucleosides to form dimers, trimers, and so forth.

Fig. 48: Deprotection, coupling, and oxidation steps of a solid phase DNA synthesis method.

Fig. 49: An illustrative synthesis route for the nucleoside building blocks used in the VLSIPS™ method.

Fig. 50: A preferred photoremovable protecting group, MeNPOC, and preparation of the group in active form.

Fig. 51: Detection system for scanning a DNA chip.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides a number of strategies for comparing a polynucleotide of known sequence (a reference sequence) with variants of that sequence (target sequences).

5 The comparison can be performed at the level of entire genomes, chromosomes, genes, exons or introns, or can focus on individual mutant sites and immediately adjacent bases. The strategies allow detection of variations, such as mutations or polymorphisms, in the target sequence irrespective whether a particular variant has previously been characterized. The
10 strategies both define the nature of a variant and identify its location in a target sequence.

The strategies employ arrays of oligonucleotide probes immobilized to a solid support. Target sequences are analyzed
15 by determining the extent of hybridization at particular probes in the array. The strategy in selection of probes facilitates distinction between perfectly matched probes and probes showing single-base or other degrees of mismatches. The strategy usually entails sampling each nucleotide of
20 interest in a target sequence several times, thereby achieving a high degree of confidence in its identity. This level of confidence is further increased by sampling of adjacent nucleotides in the target sequence to nucleotides of interest. The number of probes on the chip can be quite large (e.g.,
25 10^5 - 10^6). However, usually only a small proportion of the total number of probes of a given length are represented. Some advantage of the use of only a small proportion of all possible probes of a given length include: (i) each position in the array is highly informative, whether or not
30 hybridization occurs; (ii) nonspecific hybridization is minimized; (iii) it is straightforward to correlate hybridization differences with sequence differences, particularly with reference to the hybridization pattern of a known standard; and (iv) the ability to address each probe
35 independently during synthesis, using high resolution photolithography, allows the array to be designed and optimized for any sequence. For example the length of any probe can be varied independently of the others.

The present tiling strategies result in sequencing and comparison methods suitable for routine large-scale practice with a high degree of confidence in the sequence output.

5 I. GENERAL TILING STRATEGIES

A. Selection of Reference Sequence

The chips are designed to contain probes exhibiting complementarity to one or more selected reference sequence whose sequence is known. The chips are used to read a target
10 sequence comprising either the reference sequence itself or variants of that sequence. Target sequences may differ from the reference sequence at one or more positions but show a high overall degree of sequence identity with the reference sequence (e.g., at least 75, 90, 95, 99, 99.9 or 99.99%). Any
15 polynucleotide of known sequence can be selected as a reference sequence. Reference sequences of interest include sequences known to include mutations or polymorphisms associated with phenotypic changes having clinical significance in human patients. For example, the CFTR gene
20 and P53 gene in humans have been identified as the location of several mutations resulting in cystic fibrosis or cancer respectively. Other reference sequences of interest include those that serve to identify pathogenic microorganisms and/or are the site of mutations by which such microorganisms acquire
25 drug resistance (e.g., the HIV reverse transcriptase gene). Other reference sequences of interest include regions where polymorphic variations are known to occur (e.g., the D-loop region of mitochondrial DNA). These reference sequences have utility for, e.g., forensic or epidemiological studies. Other
30 reference sequences of interest include p34 (related to p53), p65 (implicated in breast, prostate and liver cancer), and DNA segments encoding cytochromes P450 (see Meyer et al., *Pharmac. Ther.* 46, 349-355 (1990)). Other reference sequences of
interest include those from the genome of pathogenic viruses
35 (e.g., hepatitis (A, B, or C), herpes virus (e.g., VZV, HSV-1, HAV-6, HSV-II, and CMV, Epstein Barr virus), adenovirus, influenza virus, flaviviruses, echovirus, rhinovirus, coxsackie virus, cornovirus, respiratory syncytial virus,

mumps virus, rotavirus, measles virus, rubella virus, parvovirus, vaccinia virus, HTLV virus, dengue virus, papillomavirus, molluscum virus, poliovirus, rabies virus, JC virus and arboviral encephalitis virus. Other reference sequences of interest are from genomes or episomes of pathogenic bacteria, particularly regions that confer drug resistance or allow phylogenetic characterization of the host (e.g., 16S rRNA or corresponding DNA). For example, such bacteria include chlamydia, rickettsial bacteria, mycobacteria, staphylococci, streptococci, pneumococci, meningococci and gonococci, klebsiella, proteus, serratia, pseudomonas, legionella, diphtheria, salmonella, bacilli, cholera, tetanus, botulism, anthrax, plague, leptospirosis, and Lyme disease bacteria. Other reference sequences of interest include those in which mutations result in the following autosomal recessive disorders: sickle cell anemia, β -thalassemia, phenylketonuria, galactosemia, Wilson's disease, hemochromatosis, severe combined immunodeficiency, alpha-1-antitrypsin deficiency, albinism, alcaptonuria, lysosomal storage diseases and Ehlers-Danlos syndrome. Other reference sequences of interest include those in which mutations result in X-linked recessive disorders: hemophilia, glucose-6-phosphate dehydrogenase, agammaglobulinemia, diabetes insipidus, Lesch-Nyhan syndrome, muscular dystrophy, Wiskott-Aldrich syndrome, Fabry's disease and fragile X-syndrome. Other reference sequences of interest includes those in which mutations result in the following autosomal dominant disorders: familial hypercholesterolemia, polycystic kidney disease, Huntington's disease, hereditary spherocytosis, Marfan's syndrome, von Willebrand's disease, neurofibromatosis, tuberous sclerosis, hereditary hemorrhagic telangiectasia, familial colonic polyposis, Ehlers-Danlos syndrome, myotonic dystrophy, muscular dystrophy, osteogenesis imperfecta, acute intermittent porphyria, and von Hippel-Lindau disease.

The length of a reference sequence can vary widely from a full-length genome, to an individual chromosome, episome, gene, component of a gene, such as an exon, intron or

regulatory sequences, to a few nucleotides. A reference sequence of between about 2, 5, 10, 20, 50, 100, 5000, 1000, 5,000 or 10,000, 20,000 or 100,000 nucleotides is common. Sometimes only particular regions of a sequence (e.g., exons of a gene) are of interest. In such situations, the particular regions can be considered as separate reference sequences or can be considered as components of a single reference sequence, as matter of arbitrary choice.

A reference sequence can be any naturally occurring, mutant, consensus or purely hypothetical sequence of nucleotides, RNA or DNA. For example, sequences can be obtained from computer data bases, publications or can be determined or conceived *de novo*. Usually, a reference sequence is selected to show a high degree of sequence identity to envisaged target sequences. Often, particularly, where a significant degree of divergence is anticipated between target sequences, more than one reference sequence is selected. Combinations of wildtype and mutant reference sequences are employed in several applications of the tiling strategy.

B. Chip Design

1. Basic Tiling Strategy

The basic tiling strategy provides an array of immobilized probes for analysis of target sequences showing a high degree of sequence identity to one or more selected reference sequences. The strategy is first illustrated for an array that is subdivided into four probe sets, although it will be apparent that in some situations, satisfactory results are obtained from only two probe sets. A first probe set comprises a plurality of probes exhibiting perfect complementarity with a selected reference sequence. The perfect complementarity usually exists throughout the length of the probe. However, probes having a segment or segments of perfect complementarity that is/are flanked by leading or trailing sequences lacking complementarity to the reference sequence can also be used. Within a segment of complementarity, each probe in the first probe set has at

least one interrogation position that corresponds to a nucleotide in the reference sequence. That is, the interrogation position is aligned with the corresponding nucleotide in the reference sequence, when the probe and reference sequence are aligned to maximize complementarity between the two. If a probe has more than one interrogation position, each corresponds with a respective nucleotide in the reference sequence. The identity of an interrogation position and corresponding nucleotide in a particular probe in the first probe set cannot be determined simply by inspection of the probe in the first set. As will become apparent, an interrogation position and corresponding nucleotide is defined by the comparative structures of probes in the first probe set and corresponding probes from additional probe sets.

In principle, a probe could have an interrogation position at each position in the segment complementary to the reference sequence. Sometimes, interrogation positions provide more accurate data when located away from the ends of a segment of complementarity. Thus, typically a probe having a segment of complementarity of length x does not contain more than $x-2$ interrogation positions. Since probes are typically 9-21 nucleotides, and usually all of a probe is complementary, a probe typically has 1-19 interrogation positions. Often the probes contain a single interrogation position, at or near the center of probe.

For each probe in the first set, there are, for purposes of the present illustration, three corresponding probes from three additional probe sets. See Fig. 1. Thus, there are four probes corresponding to each nucleotide of interest in the reference sequence. Each of the four corresponding probes has an interrogation position aligned with that nucleotide of interest. Usually, the probes from the three additional probe sets are identical to the corresponding probe from the first probe set with one exception. The exception is that at least one (and often only one) interrogation position, which occurs in the same position in each of the four corresponding probes from the four probe sets, is occupied by a different nucleotide in the four probe sets. For example, for an A

nucleotide in the reference sequence, the corresponding probe from the first probe set has its interrogation position occupied by a T, and the corresponding probes from the additional three probe sets have their respective
5 interrogation positions occupied by A, C, or G, a different nucleotide in each probe. Of course, if a probe from the first probe set comprises trailing or flanking sequences lacking complementarity to the reference sequences (see Fig. 2), these sequences need not be present in corresponding
10 probes from the three additional sets. Likewise corresponding probes from the three additional sets can contain leading or trailing sequences outside the segment of complementarity that are not present in the corresponding probe from the first probe set. Occasionally, the probes from the additional three
15 probe set are identical (with the exception of interrogation position(s)) to a contiguous subsequence of the full complementary segment of the corresponding probe from the first probe set. In this case, the subsequence includes the interrogation position and usually differs from the full-
20 length probe only in the omission of one or both terminal nucleotides from the termini of a segment of complementarity. That is, if a probe from the first probe set has a segment of complementarity of length n , corresponding probes from the other sets will usually include a subsequence of the segment
25 of at least length $n-2$. Thus, the subsequence is usually at least 3, 4, 7, 9, 15, 21, or 25 nucleotides long, most typically, in the range of 9-21 nucleotides. The subsequence should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference
30 sequence mutated at the interrogation position than to the reference sequence.

The probes can be oligodeoxyribonucleotides or oligoribonucleotides, or any modified forms of these polymers that are capable of hybridizing with a target nucleic sequence
35 by complementary base-pairing. Complementary base pairing means sequence-specific base pairing which includes e.g., Watson-Crick base pairing as well as other forms of base pairing such as Hoogsteen base pairing. Modified forms

include 2'-O-methyl oligoribonucleotides and so-called PNAs, in which oligodeoxyribonucleotides are linked via peptide bonds rather than phosphodiester bonds. The probes can be attached by any linkage to a support (e.g., 3', 5' or via the base). 3' attachment is more usual as this orientation is compatible with the preferred chemistry for solid phase synthesis of oligonucleotides.

The number of probes in the first probe set (and as a consequence the number of probes in additional probe sets) depends on the length of the reference sequence, the number of nucleotides of interest in the reference sequence and the number of interrogation positions per probe. In general, each nucleotide of interest in the reference sequence requires the same interrogation position in the four sets of probes.

Consider, as an example, a reference sequence of 100 nucleotides, 50 of which are of interest, and probes each having a single interrogation position. In this situation, the first probe set requires fifty probes, each having one interrogation position corresponding to a nucleotide of interest in the reference sequence. The second, third and fourth probe sets each have a corresponding probe for each probe in the first probe set, and so each also contains a total of fifty probes. The identity of each nucleotide of interest in the reference sequence is determined by comparing the relative hybridization signals at four probes having interrogation positions corresponding to that nucleotide from the four probe sets.

In some reference sequences, every nucleotide is of interest. In other reference sequences, only certain portions in which variants (e.g., mutations or polymorphisms) are concentrated are of interest. In other reference sequences, only particular mutations or polymorphisms and immediately adjacent nucleotides are of interest. Usually, the first probe set has interrogation positions selected to correspond to at least a nucleotide (e.g., representing a point mutation) and one immediately adjacent nucleotide. Usually, the probes in the first set have interrogation positions corresponding to at least 3, 10, 50, 100, 1000, or 20,000 contiguous

nucleotides. The probes usually have interrogation positions corresponding to at least 5, 10, 30, 50, 75, 90, 99 or sometimes 100% of the nucleotides in a reference sequence. Frequently, the probes in the first probe set completely span the reference sequence and overlap with one another relative to the reference sequence. For example, in one common arrangement each probe in the first probe set differs from another probe in that set by the omission of a 3' base complementary to the reference sequence and the acquisition of a 5' base complementary to the reference sequence. See Fig. 3.

For conceptual simplicity, the probes in a set are usually arranged in order of the sequence in a lane across the chip. A lane contains a series of overlapping probes, which represent or tile across, the selected reference sequence (see Fig. 3). The components of the four sets of probes are usually laid down in four parallel lanes, collectively constituting a row in the horizontal direction and a series of 4-member columns in the vertical direction. Corresponding probes from the four probe sets (i.e., complementary to the same subsequence of the reference sequence) occupy a column. Each probe in a lane usually differs from its predecessor in the lane by the omission of a base at one end and the inclusion of additional base at the other end as shown in Fig. 3. However, this orderly progression of probes can be interrupted by the inclusion of control probes or omission of probes in certain columns of the array. Such columns serve as controls to orient the chip, or gauge the background, which can include target sequence nonspecifically bound to the chip.

The probes sets are usually laid down in lanes such that all probes having an interrogation position occupied by an A form an A-lane, all probes having an interrogation position occupied by a C form a C-lane, all probes having an interrogation position occupied by a G form a G-lane, and all probes having an interrogation position occupied by a T (or U) form a T lane (or a U lane). Note that in this arrangement there is not a unique correspondence between probe sets and lanes. Thus, the probe from the first probe set is laid down

in the A-lane, C-lane, A-lane, A-lane and T-lane for the five columns in Fig. 4. The interrogation position on a column of probes corresponds to the position in the target sequence whose identity is determined from analysis of hybridization to the probes in that column. Thus, I_1 - I_5 respectively correspond to N_1 - N_5 in Fig. 4. The interrogation position can be anywhere in a probe but is usually at or near the central position of the probe to maximize differential hybridization signals between a perfect match and a single-base mismatch. For example, for an 11 mer probe, the central position is the sixth nucleotide.

Although the array of probes is usually laid down in rows and columns as described above, such a physical arrangement of probes on the chip is not essential. Provided that the spatial location of each probe in an array is known, the data from the probes can be collected and processed to yield the sequence of a target irrespective of the physical arrangement of the probes on a chip. In processing the data, the hybridization signals from the respective probes can be reassorted into any conceptual array desired for subsequent data reduction whatever the physical arrangement of probes on the chip.

A range of lengths of probes can be employed in the chips. As noted above, a probe may consist exclusively of a complementary segments, or may have one or more complementary segments juxtaposed by flanking, trailing and/or intervening segments. In the latter situation, the total length of complementary segment(s) is more important than the length of the probe. In functional terms, the complementarity segment(s) of the first probe sets should be sufficiently long to allow the probe to hybridize detectably more strongly to a reference sequence compared with a variant of the reference including a single base mutation at the nucleotide corresponding to the interrogation position of the probe. Similarly, the complementarity segment(s) in corresponding probes from additional probe sets should be sufficiently long to allow a probe to hybridize detectably more strongly to a variant of the reference sequence having a single nucleotide

substitution at the interrogation position relative to the reference sequence. A probe usually has a single complementary segment having a length of at least 3 nucleotides, and more usually at least 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 or 30 bases exhibiting perfect complementarity (other than possibly at the interrogation position(s) depending on the probe set) to the reference sequence. In bridging strategies, where more than one segment of complementarity is present, each segment provides at least three complementary nucleotides to the reference sequence and the combined segments provide at least two segments of three or a total of six complementary nucleotides. As in the other strategies, the combined length of complementary segments is typically from 6-30 nucleotides, and preferably from about 9-21 nucleotides. The two segments are often approximately the same length. Often, the probes (or segment of complementarity within probes) have an odd number of bases, so that an interrogation position can occur in the exact center of the probe.

In some chips, all probes are the same length. Other chips employ different groups of probe sets, in which case the probes are of the same size within a group, but differ between different groups. For example, some chips have one group comprising four sets of probes as described above in which all the probes are 11 mers, together with a second group comprising four sets of probes in which all of the probes are 13 mers. Of course, additional groups of probes can be added. Thus, some chips contain, e.g., four groups of probes having sizes of 11 mers, 13 mers, 15 mers and 17 mers. Other chips have different size probes within the same group of four probe sets. In these chips, the probes in the first set can vary in length independently of each other. Probes in the other sets are usually the same length as the probe occupying the same column from the first set. However, occasionally different lengths of probes can be included at the same column position in the four lanes. The different length probes are included to equalize hybridization signals from probes irrespective of

whether A-T or C-G bonds are formed at the interrogation position.

The length of probe can be important in distinguishing between a perfectly matched probe and probes showing a single-base mismatch with the target sequence. The discrimination is usually greater for short probes. Shorter probes are usually also less susceptible to formation of secondary structures. However, the absolute amount of target sequence bound, and hence the signal, is greater for larger probes. The probe length representing the optimum compromise between these competing considerations may vary depending on *inter alia* the GC content of a particular region of the target DNA sequence, secondary structure, synthesis efficiency and cross-hybridization. In some regions of the target, depending on hybridization conditions, short probes (e.g., 11 mers) may provide information that is inaccessible from longer probes (e.g., 19 mers) and vice versa. Maximum sequence information can be read by including several groups of different sized probes on the chip as noted above. However, for many regions of the target sequence, such a strategy provides redundant information in that the same sequence is read multiple times from the different groups of probes. Equivalent information can be obtained from a single group of different sized probes in which the sizes are selected to maximize readable sequence at particular regions of the target sequence. The appropriate size of probes at different regions of the target sequence can be determined from, e.g., Fig. 12, which compares the readability of different sized probes in different regions of a target. The strategy of customizing probe length within a single group of probe sets minimizes the total number of probes required to read a particular target sequence. This leaves ample capacity for the chip to include probes to other reference sequences.

The invention provides an optimization block which allows systematic variation of probe length and interrogation position to optimize the selection of probes for analyzing a particular nucleotide in a reference sequence. The block comprises alternating columns of probes complementary to the

wildtype target and probes complementary to a specific mutation. The interrogation position is varied between columns and probe length is varied down a column.

Hybridization of the chip to the reference sequence or the mutant form of the reference sequence identifies the probe length and interrogation position providing the greatest differential hybridization signal.

The probes are designed to be complementary to either strand of the reference sequence (e.g., coding or non-coding).

Some chips contain separate groups of probes, one complementary to the coding strand, the other complementary to the noncoding strand. Independent analysis of coding and noncoding strands provides largely redundant information. However, the regions of ambiguity in reading the coding strand are not always the same as those in reading the noncoding strand. Thus, combination of the information from coding and noncoding strands increases the overall accuracy of sequencing.

Some chips contain additional probes or groups of probes designed to be complementary to a second reference sequence. The second reference sequence is often a subsequence of the first reference sequence bearing one or more commonly occurring mutations or interstrain variations. The second group of probes is designed by the same principles as described above except that the probes exhibit complementarity to the second reference sequence. The inclusion of a second group is particularly useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (i.e., two or more mutations within 9 to 21 bases). Of course, the same principle can be extended to provide chips containing groups of probes for any number of reference sequences. Alternatively, the chips may contain additional probe(s) that do not form part of a tiled array as noted above, but rather serves as probe(s) for a conventional reverse dot blot. For example, the presence of mutation can be detected from binding of a target sequence to a single oligomeric probe harboring the mutation. Preferably, an

additional probe containing the equivalent region of the wildtype sequence is included as a control.

The chips are read by comparing the intensities of labelled target bound to the probes in an array.

5 Specifically, a comparison is performed between each lane of probes (e.g., A, C, G and T lanes) at each columnar position (physical or conceptual). For a particular columnar position, the lane showing the greatest hybridization signal is called as the nucleotide present at the position in the target
10 sequence corresponding to the interrogation position in the probes. See Fig. 5. The corresponding position in the target sequence is that aligned with the interrogation position in corresponding probes when the probes and target are aligned to maximize complementarity. Of the four probes in a column,
15 only one can exhibit a perfect match to the target sequence whereas the others usually exhibit at least a one base pair mismatch. The probe exhibiting a perfect match usually produces a substantially greater hybridization signal than the other three probes in the column and is thereby easily
20 identified. However, in some regions of the target sequence, the distinction between a perfect match and a one-base mismatch is less clear. Thus, a call ratio is established to define the ratio of signal from the best hybridizing probes to the second best hybridizing probe that must be exceeded for a
25 particular target position to be read from the probes. A high call ratio ensures that few if any errors are made in calling target nucleotides, but can result in some nucleotides being scored as ambiguous, which could in fact be accurately read. A lower call ratio results in fewer ambiguous calls, but can
30 result in more erroneous calls. It has been found that at a call ratio of 1.2 virtually all calls are accurate. However, a small but significant number of bases (e.g., up to about 10%) may have to be scored as ambiguous.

Although small regions of the target sequence can
35 sometimes be ambiguous, these regions usually occur at the same or similar segments in different target sequences. Thus, for precharacterized mutations, it is known in advance whether

that mutation is likely to occur within a region of unambiguously determinable sequence.

An array of probes is most useful for analyzing the reference sequence from which the probes were designed and variants of that sequence exhibiting substantial sequence similarity with the reference sequence (e.g., several single-base mutants spaced over the reference sequence). When an array is used to analyze the exact reference sequence from which it was designed, one probe exhibits a perfect match to the reference sequence, and the other three probes in the same column exhibits single-base mismatches. Thus, discrimination between hybridization signals is usually high and accurate sequence is obtained. High accuracy is also obtained when an array is used for analyzing a target sequence comprising a variant of the reference sequence that has a single mutation relative to the reference sequence, or several widely spaced mutations relative to the reference sequence. At different mutant loci, one probe exhibits a perfect match to the target, and the other three probes occupying the same column exhibit single-base mismatches, the difference (with respect to analysis of the reference sequence) being the lane in which the perfect match occurs.

For target sequences showing a high degree of divergence from the reference strain or incorporating several closely spaced mutations from the reference strain, a single group of probes (i.e., designed with respect to a single reference sequence) will not always provide accurate sequence for the highly variant region of this sequence. At some particular columnar positions, it may be that no single probe exhibits perfect complementarity to the target and that any comparison must be based on different degrees of mismatch between the four probes. Such a comparison does not always allow the target nucleotide corresponding to that columnar position to be called. Deletions in target sequences can be detected by loss of signal from probes having interrogation positions encompassed by the deletion. However, signal may also be lost from probes having interrogation positions closely proximal to the deletion resulting in some regions of the target sequence

that cannot be read. Target sequence bearing insertions will also exhibit short regions including and proximal to the insertion that usually cannot be read.

The presence of short regions of difficult-to-read target
5 because of closely spaced mutations, insertions or deletion,
does not prevent determination of the remaining sequence of
the target as different regions of a target sequence are
determined independently. Moreover, such ambiguities as might
result from analysis of diverse variants with a single group
10 of probes can be avoided by including multiple groups of probe
sets on a chip. For example, one group of probes can be
designed based on a full-length reference sequence, and the
other groups on subsequences of the reference sequence
incorporating frequently occurring mutations or strain
15 variations.

A particular advantage of the present sequencing strategy
over conventional sequencing methods is the capacity
simultaneously to detect and quantify proportions of multiple
target sequences. Such capacity is valuable, e.g., for
20 diagnosis of patients who are heterozygous with respect to a
gene or who are infected with a virus, such as HIV, which is
usually present in several polymorphic forms. Such capacity
is also useful in analyzing targets from biopsies of tumor
cells and surrounding tissues. The presence of multiple
25 target sequences is detected from the relative signals of the
four probes at the array columns corresponding to the target
nucleotides at which diversity occurs. The relative signals
at the four probes for the mixture under test are compared
with the corresponding signals from a homogeneous reference
30 sequence. An increase in a signal from a probe that is
mismatched with respect to the reference sequence, and a
corresponding decrease in the signal from the probe which is
matched with the reference sequence signal the presence of a
mutant strain in the mixture. The extent in shift in
35 hybridization signals of the probes is related to the
proportion of a target sequence in the mixture. Shifts in
relative hybridization signals can be quantitatively related
to proportions of reference and mutant sequence by prior

calibration of the chip with seeded mixtures of the mutant and reference sequences. By this means, a chip can be used to detect variant or mutant strains constituting as little as 1, 5, 20, or 25 % of a mixture of stains.

5 Similar principles allow the simultaneous analysis of multiple target sequences even when none is identical to the reference sequence. For example, with a mixture of two target sequences bearing first and second mutations, there would be a variation in the hybridization patterns of probes having
10 interrogation positions corresponding to the first and second mutations relative to the hybridization pattern with the reference sequence. At each position, one of the probes having a mismatched interrogation position relative to the reference sequence would show an increase in hybridization
15 signal, and the probe having a matched interrogation position relative to the reference sequence would show a decrease in hybridization signal. Analysis of the hybridization pattern of the mixture of mutant target sequences, preferably in comparison with the hybridization pattern of the reference
20 sequence, indicates the presence of two mutant target sequences, the position and nature of the mutation in each strain, and the relative proportions of each strain.

 In a variation of the above method, the different components in a mixture of target sequences are differentially
25 labelled before being applied to the array. For example, a variety of fluorescent labels emitting at different wavelength are available. The use of differential labels allows independent analysis of different targets bound simultaneously to the array. For example, the methods permit comparison of
30 target sequences obtained from a patient at different stages of a disease.

2. Omission of Probes

 The general strategy outlined above employs four probes
35 to read each nucleotide of interest in a target sequence. One probe (from the first probe set) shows a perfect match to the reference sequence and the other three probes (from the second, third and fourth probe sets) exhibit a mismatch with

the reference sequence and a perfect match with a target sequence bearing a mutation at the nucleotide of interest. The provision of three probes from the second, third and fourth probe sets allows detection of each of the three possible nucleotide substitutions of any nucleotide of interest. However, in some reference sequences or regions of reference sequences, it is known in advance that only certain mutations are likely to occur. Thus, for example, at one site it might be known that an A nucleotide in the reference sequence may exist as a T mutant in some target sequences but is unlikely to exist as a C or G mutant. Accordingly, for analysis of this region of the reference sequence, one might include only the first and second probe sets, the first probe set exhibiting perfect complementarity to the reference sequence, and the second probe set having an interrogation position occupied by an invariant A residue (for detecting the T mutant). In other situations, one might include the first, second and third probes sets (but not the fourth) for detection of a wildtype nucleotide in the reference sequence and two mutant variants thereof in target sequences. In some chips, probes that would detect silent mutations (i.e., not affecting amino acid sequence) are omitted.

In some chips, the probes from the first probe set are omitted corresponding to some or all positions of the reference sequences. Such chips comprise at least two probe sets. The first probe set has a plurality of probes. Each probe comprises a segment exactly complementary to a subsequence of a reference sequence except in at least one interrogation position. A second probe set has a corresponding probe for each probe in the first probe set. The corresponding probe in the second probe set is identical to a sequence comprising the corresponding probe from the first probe set or a subsequence thereof that includes the at least one (and usually only one) interrogation position except that the at least one interrogation position is occupied by a different nucleotide in each of the two corresponding probes from the first and second probe sets. A third probe set, if present, also comprises a corresponding probe for each probe

in the first probe set except at the at least one interrogation position, which differs in the corresponding probes from the three sets. Omission of probes having a segment exhibiting perfect complementarity to the reference sequence results in loss of control information, i.e., the detection of nucleotides in a target sequence that are the same as those in a reference sequence. However, similar information can be obtained by hybridizing a chip lacking probes from the first probe set to both target and reference sequences. The hybridization can be performed sequentially, or concurrently, if the target and reference are differentially labelled. In this situation, the presence of a mutation is detected by a shift in the background hybridization intensity of the reference sequence to a perfectly matched hybridization signal of the target sequence, rather than by a comparison of the hybridization intensities of probes from the first set with corresponding probes from the second, third and fourth sets.

3. Wildtype Probe Lane

When the chips comprise four probe sets, as discussed *supra*, and the probe sets are laid down in four lanes, an A lane, a C-lane, a G lane and a T or U lane, the probe having a segment exhibiting perfect complementarity to a reference sequence varies between the four lanes from one column to another. This does not present any significant difficulty in computer analysis of the data from the chip. However, visual inspection of the hybridization pattern of the chip is sometimes facilitated by provision of an extra lane of probes, in which each probe has a segment exhibiting perfect complementarity to the reference sequence. See Fig. 4. This segment is identical to a segment from one of the probes in the other four lanes (which lane depending on the column position). The extra lane of probes (designated the wildtype lane) hybridizes to a target sequence at all nucleotide positions except those in which deviations from the reference sequence occurs. The hybridization pattern of the wildtype lane thereby provides a simple visual indication of mutations.

4. Deletion, Insertion and Multiple-Mutation Probes

Some chips provide an additional probe set specifically designed for analyzing deletion mutations. The additional probe set comprises a probe corresponding to each probe in the first probe set as described above. However, a probe from the additional probe set differs from the corresponding probe in the first probe set in that the nucleotide occupying the interrogation position is deleted in the probe from the additional probe set. See Fig. 6. Optionally, the probe from the additional probe set bears an additional nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set will hybridize more strongly than the corresponding probe from the first probe set to a target sequence having a single base deletion at the nucleotide corresponding to the interrogation position. Additional probe sets are provided in which not only the interrogation position, but also an adjacent nucleotide is detected.

Similarly, other chips provide additional probe sets for analyzing insertions. For example, one additional probe set has a probe corresponding to each probe in the first probe set as described above. However, the probe in the additional probe set has an extra T nucleotide inserted adjacent to the interrogation position. See Fig. 6. Optionally, the probe has one fewer nucleotide at one of its termini relative to the corresponding probe from the first probe set. The probe from the additional probe set hybridizes more strongly than the corresponding probe from the first probe set to a target sequence having an A nucleotide inserted in a position adjacent to that corresponding to the interrogation position. Similar additional probe sets are constructed having C, G or T/U nucleotides inserted adjacent to the interrogation position. Usually, four such probe sets, one for each nucleotide, are used in combination.

Other chips provide additional probes (multiple-mutation probes) for analyzing target sequences having multiple closely spaced mutations. A multiple-mutation probe is usually identical to a corresponding probe from the first set as

described above, except in the base occupying the interrogation position, and except at one or more additional positions, corresponding to nucleotides in which substitution may occur in the reference sequence. The one or more additional positions in the multiple mutation probe are occupied by nucleotides complementary to the nucleotides occupying corresponding positions in the reference sequence when the possible substitutions have occurred.

5. Block Tiling

As noted in the discussion of the general tiling strategy, a probe in the first probe set sometimes has more than one interrogation position. In this situation, a probe in the first probe set is sometimes matched with multiple groups of at least one, and usually, three additional probe sets. See Fig. 7. Three additional probe sets are used to allow detection of the three possible nucleotide substitutions at any one position. If only certain types of substitution are likely to occur (e.g., transitions), only one or two additional probe sets are required (analogous to the use of probes in the basic tiling strategy). To illustrate for the situation where a group comprises three additional probe sets, a first such group comprises second, third and fourth probe sets, each of which has a probe corresponding to each probe in the first probe set. The corresponding probes from the second, third and fourth probes sets differ from the corresponding probe in the first set at a first of the interrogation positions. Thus, the relative hybridization signals from corresponding probes from the first, second, third and fourth probe sets indicate the identity of the nucleotide in a target sequence corresponding to the first interrogation position. A second group of three probe sets (designated fifth, sixth and seventh probe sets), each also have a probe corresponding to each probe in the first probe set. These corresponding probes differ from that in the first probe set at a second interrogation position. The relative hybridization signals from corresponding probes from the first, fifth, sixth, and seventh probe sets indicate the identity of the nucleotide in the target sequence

corresponding to the second interrogation position. As noted above, the probes in the first probe set often have seven or more interrogation positions. If there are seven interrogation positions, there are seven groups of three additional probe sets, each group of three probe sets serving to identify the nucleotide corresponding to one of the seven interrogation positions.

Each block of probes allows short regions of a target sequence to be read. For example, for a block of probes having seven interrogation positions, seven nucleotides in the target sequence can be read. Of course, a chip can contain any number of blocks depending on how many nucleotides of the target are of interest. The hybridization signals for each block can be analyzed independently of any other block. The block tiling strategy can also be combined with other tiling strategies, with different parts of the same reference sequence being tiled by different strategies.

The block tiling strategy offers two advantages over the basic strategy in which each probe in the first set has a single interrogation position. One advantage is that the same sequence information can be obtained from fewer probes. A second advantage is that each of the probes constituting a block (i.e., a probe from the first probe set and a corresponding probe from each of the other probe sets) can have identical 3' and 5' sequences, with the variation confined to a central segment containing the interrogation positions. The identity of 3' sequence between different probes simplifies the strategy for solid phase synthesis of the probes on the chip and results in more uniform deposition of the different probes on the chip, thereby in turn increasing the uniformity of signal to noise ratio for different regions of the chip. A third advantage is that greater signal uniformity is achieved within a block.

35 6. Multiplex Tiling

In the block tiling strategy discussed above, the identity of a nucleotide in a target or reference sequence is determined by comparison of hybridization patterns of one

probe having a segment showing a perfect match with that of other probes (usually three other probes) showing a single base mismatch. In multiplex tiling, the identity of at least two nucleotides in a reference or target sequence is

5 determined by comparison of hybridization signal intensities of four probes, two of which have a segment showing perfect complementarity or a single base mismatch to the reference sequence, and two of which have a segment showing perfect complementarity or a double-base mismatch to a segment. The

10 four probes whose hybridization patterns are to be compared each have a segment that is exactly complementary to a reference sequence except at two interrogation positions, in which the segment may or may not be complementary to the reference sequence. The interrogation positions correspond to

15 the nucleotides in a reference or target sequence which are determined by the comparison of intensities. The nucleotides occupying the interrogation positions in the four probes are selected according to the following rule. The first interrogation position is occupied by a different nucleotide

20 in each of the four probes. The second interrogation position is also occupied by a different nucleotide in each of the four probes. In two of the four probes, designated the first and second probes, the segment is exactly complementary to the reference sequence except at not more than one of the two

25 interrogation positions. In other words, one of the interrogation positions is occupied by a nucleotide that is complementary to the corresponding nucleotide from the reference sequence and the other interrogation position may or may not be so occupied. In the other two of the four probes,

30 designated the third and fourth probes, the segment is exactly complementary to the reference sequence except that both interrogation positions are occupied by nucleotides which are noncomplementary to the respective corresponding nucleotides in the reference sequence.

35 There are number of ways of satisfying these conditions depending on whether the two nucleotides in the reference sequence corresponding to the two interrogation positions are the same or different. If these two nucleotides are different

in the reference sequence (probability $3/4$), the conditions are satisfied by each of the two interrogation positions being occupied by the same nucleotide in any given probe. For example, in the first probe, the two interrogation positions
5 would both be A, in the second probe, both would be C, in the third probe, each would be G, and in the fourth probe each would be T or U. If the two nucleotides in the reference sequence corresponding to the two interrogation positions are different, the conditions noted above are satisfied by each of
10 the interrogation positions in any one of the four probes being occupied by complementary nucleotides. For example, in the first probe, the interrogation positions could be occupied by A and T, in the second probe by C and G, in the third probe by G and C, and in the fourth probe, by T and A. See (Fig. 8).

15 When the four probes are hybridized to a target that is the same as the reference sequence or differs from the reference sequence at one (but not both) of the interrogation positions, two of the four probes show a double-mismatch with the target and two probes show a single mismatch. The
20 identity of probes showing these different degrees of mismatch can be determined from the different hybridization signals. From the identity of the probes showing the different degrees of mismatch, the nucleotides occupying both of the interrogation positions in the target sequence can be deduced.

25 For ease of illustration, the multiplex strategy has been initially described for the situation where there are two nucleotides of interest in a reference sequence and only four probes in an array. Of course, the strategy can be extended to analyze any number of nucleotides in a target sequence by
30 using additional probes. In one variation, each pair of interrogation positions is read from a unique group of four probes. In a block variation, different groups of four probes exhibit the same segment of complementarity with the reference sequence, but the interrogation positions move within a block.
35 The block and standard multiplex tiling variants can of course be used in combination for different regions of a reference sequence. Either or both variants can also be used in combination with any of the other tiling strategies described.

7. Helper Mutations

Occasionally small regions of a reference sequence give a low hybridization signal as a result of annealing of probes. The self-annealing reduces the amount of probe effectively available for hybridizing to the target. Although such regions of the target are generally small and the reduction of hybridization signal is usually not so substantial as to obscure the sequence of this region, this concern can be avoided by the use of probes incorporating helper mutations. The helper mutation(s) serve to break-up regions of internal complementarity within a probe and thereby prevent annealing. Usually, one or two helper mutations are quite sufficient for this purpose. The inclusion of helper mutations can be beneficial in any of the tiling strategies noted above. In general each probe having a particular interrogation position has the same helper mutation(s). Thus, such probes have a segment in common which shows perfect complementarity with a reference sequence, except that the segment contains at least one helper mutation (the same in each of the probes) and at least one interrogation position (different in all of the probes). For example, in the basic tiling strategy, a probe from the first probe set comprises a segment containing an interrogation position and showing perfect complementarity with a reference sequence except for one or two helper mutations. The corresponding probes from the second, third and fourth probe sets usually comprise the same segment (or sometimes a subsequence thereof including the helper mutation(s) and interrogation position), except that the base occupying the interrogation position varies in each probe. See Fig. 9.

Usually, the helper mutation tiling strategy is used in conjunction with one of the tiling strategies described above. The probes containing helper mutations are used to tile regions of a reference sequence otherwise giving low hybridization signal (e.g., because of self-complementarity), and the alternative tiling strategy is used to tile intervening regions.

8. Pooling Strategies

Pooling strategies also employ arrays of immobilized probes. Probes are immobilized in cells of an array, and the hybridization signal of each cell can be determined independently of any other cell. A particular cell may be occupied by pooled mixture of probes. Although the identity of each probe in the mixture is known, the individual probes in the pool are not separately addressable. Thus, the hybridization signal from a cell is the aggregate of that of the different probes occupying the cell. In general, a cell is scored as hybridizing to a target sequence if at least one probe occupying the cell comprises a segment exhibiting perfect complementarity to the target sequence.

A simple strategy to show the increased power of pooled strategies over a standard tiling is to create three cells each containing a pooled probe having a single pooled position, the pooled position being the same in each of the pooled probes. At the pooled position, there are two possible nucleotide, allowing the pooled probe to hybridize to two target sequences. In tiling terminology, the pooled position of each probe is an interrogation position. As will become apparent, comparison of the hybridization intensities of the pooled probes from the three cells reveals the identity of the nucleotide in the target sequence corresponding to the interrogation position (i.e., that is matched with the interrogation position when the target sequence and pooled probes are maximally aligned for complementarity).

The three cells are assigned probe pools that are perfectly complementary to the target except at the pooled position, which is occupied by a different pooled nucleotide in each probe as follows:

[AC] = M, [GT]=K, [AG]=R
 as substitutions in the probe
 IUPAC standard ambiguity notation)

X - interrogation position

5 Target: TAACCACTCACGGGAGCA

Pool 1: ATTGGMGAGTGCCC
 =ATTGGaGAGTGCCC (complement to mutant 't')
 +ATTGGcGAGTGCCC (complement to mutant 'g')

10 Pool 2: ATTGGKGAGTGCCC
 =ATTGGgGAGTGCCC (complement to mutant 'c')
 +ATTGGtGAGTGCCC (complement to wild type 'a')

15 Pool 3: ATTGGRGAGTGCCC
 =ATTGGaGAGTGCCC (complement to mutant 't')
 +ATTGGgGAGTGCCC (complement to mutant 'c')

20 With 3 pooled probes, all 4 possible single base pair states
 (wild and 3 mutants) are detected. A pool hybridizes with a
 target if some probe contained within that pool is
 complementary to that target.

25		Hybridization?
	Pool:	1 2 3
	Target: TAACCACTCACGGGAGCA	n y n
	Mutant: TAACCCCTCACGGGAGCA	n y y
	Mutant: TAACCgCTCACGGGAGCA	y n n
30	Mutant: TAACCTCTCACGGGAGCA	y n y

A cell containing a pair (or more) of oligonucleotides
 lights up when a target complementary to any of the
 oligonucleotide in the cell is present. Using the simple
 35 strategy, each of the four possible targets (wild and three
 mutants) yields a unique hybridization pattern among the three
 cells.

Since a different pattern of hybridizing pools is
 obtained for each possible nucleotide in the target sequence
 40 corresponding to the pooled interrogation position in the
 probes, the identity of the nucleotide can be determined from
 the hybridization pattern of the pools. Whereas, a standard
 tiling requires four cells to detect and identify the possible
 single-base substitutions at one location, this simple pooled
 45 strategy only requires three cells.

A more efficient pooling strategy for sequence analysis is the 'Trellis' strategy. In this strategy, each pooled probe has a segment of perfect complementarity to a reference sequence except at three pooled positions. One pooled position is an N pool. The three pooled positions may or may not be contiguous in a probe. The other two pooled positions are selected from the group of three pools consisting of (1) M or K, (2) R or Y and (3) W or S, where the single letters are IUPAC standard ambiguity codes. The sequence of a pooled probe is thus, of the form XXXN[(M/K) or (R/Y) or (W/S)][(M/K) or (R/Y) or (W/S)]XXXXX, where XXX represents bases complementary to the reference sequence. The three pooled positions may be in any order, and may be contiguous or separated by intervening nucleotides. For, the two positions occupied by [(M/K) or (R/Y) or (W/S)], two choices must be made. First, one must select one of the following three pairs of pooled nucleotides (1) M/K, (2) R/Y and (3) W/S. The one of three pooled nucleotides selected may be the same or different at the two pooled positions. Second, supposing, for example, one selects M/K at one position, one must then choose between M or K. This choice should result in selection of a pooled nucleotide comprising a nucleotide that complements the corresponding nucleotide in a reference sequence, when the probe and reference sequence are maximally aligned. The same principle governs the selection between R and Y, and between W and S. A trellis pool probe has one pooled position with four possibilities, and two pooled positions, each with two possibilities. Thus, a trellis pool probe comprises a mixture of 16 (4 x 2 x 2) probes. Since each pooled position includes one nucleotide that complements the corresponding nucleotide from the reference sequence, one of these 16 probes has a segment that is the exact complement of the reference sequence. A target sequence that is the same as the reference sequence (i.e., a wildtype target) gives a hybridization signal to each probe cell. Here, as in other tiling methods, the segment of complementarity should be sufficiently long to permit specific hybridization of a pooled probe to a reference sequence be detected relative to a variant of that reference

sequence. Typically, the segment of complementarity is about 9-21 nucleotides.

A target sequence is analyzed by comparing hybridization intensities at three pooled probes, each having the structure described above. The segments complementary to the reference sequence present in the three pooled probes show some overlap. Sometimes the segments are identical (other than at the interrogation positions). However, this need not be the case. For example, the segments can tile across a reference sequence in increments of one nucleotide (i.e., one pooled probe differs from the next by the acquisition of one nucleotide at the 5' end and loss of a nucleotide at the 3' end). The three interrogation positions may or may not occur at the same relative positions within each pooled probe (i.e., spacing from a probe terminus). All that is required is that one of the three interrogation positions from each of the three pooled probes aligns with the same nucleotide in the reference sequence, and that this interrogation position is occupied by a different pooled nucleotide in each of the three probes. In one of the three probes, the interrogation position is occupied by an N. In the other two pooled probes the interrogation position is occupied by one of (M/K) or (R/Y) or (W/S).

In the simplest form of the trellis strategy, three pooled probes are used to analyze a single nucleotide in the reference sequence. Much greater economy of probes is achieved when more pooled probes are included in an array. For example, consider an array of five pooled probes each having the general structure outlined above. Three of these pooled probes have an interrogation position that aligns with the same nucleotide in the reference sequence and are used to read that nucleotide. A different combination of three probes have an interrogation position that aligns with a different nucleotide in the reference sequence. Comparison of these three probe intensities allows analysis of this second nucleotide. Still another combination of three pooled probes from the set of five have an interrogation position that aligns with a third nucleotide in the reference sequence and

these probes are used to analyze that nucleotide. Thus, three nucleotides in the reference sequence are fully analyzed from only five pooled probes. By comparison, the basic tiling strategy would require 12 probes for a similar analysis.

5 As an example, a pooled probe for analysis of a target sequence by the trellis strategy is shown below:

Target: ATTAACCACTCACGGGAGCTCT

Pool: TGGTGKNKYGCCCT

10

The pooled probe actually comprises 16 individual probes:

15 TGGTGAGcGCCCT
+TGGTGcGcGCCCT
+TGGTGgGcGCCCT
+TGGTGtGcGCCCT
+TGGTGAtcGCCCT
+TGGTGctcGCCCT
+TGGTGgtcGCCCT
20 +TGGTGttcGCCCT
+TGGTGAGTGCCCT
+TGGTGcGTGCCCT
+TGGTGgGTGCCCT
+TGGTGtGTGCCCT
25 +TGGTGAtTGCCCT
+TGGTGctTGCCCT
+TGGTGgtTGCCCT
+TGGTGttTGCCCT

30

The trellis strategy employs an array of probes having at least three cells, each of which is occupied by a pooled probe as described above.

35 Consider the use of three such pooled probes for analyzing a target sequence, of which one position may contain any single base substitution to the reference sequence (i.e, there are four possible target sequences to be distinguished). Three cells are occupied by pooled probes having a pooled interrogation position corresponding to the position of
40 possible substitution in the target sequence, one cell with an 'N', one cell with one of 'M' or 'K', and one cell with 'R' or 'Y'. An interrogation position corresponds to a nucleotide in the target sequence if it aligns adjacent with that nucleotide when the probe and target sequence are aligned to maximize
45 complementarity. Note that although each of the pooled

probes has two other pooled positions, these positions are not relevant for the present illustration. The positions are only relevant when more than one position in the target sequence is to be read, a circumstance that will be considered later. For

5 present purposes, the cell with the 'N' in the interrogation position lights up for the wildtype sequence and any of the three single base substitutions of the target sequence. The cell with M/K in the interrogation position lights up for the wildtype sequence and one of the single-base substitutions.

10 The cell with R/Y in the interrogation position lights up for the wildtype sequence and a second of the single-base substitutions. Thus, the four possible target sequences hybridize to the three pools of probes in four distinct patterns, and the four possible target sequences can be
15 distinguished.

To illustrate further, consider four possible target sequences (differing at a single position) and a pooled probe having three pooled positions, N, K and Y with the Y position as the interrogation position (i.e., aligned with the variable
20 position in the target sequence):

Target

Wild: ATTAACCACTCACGGGAGCTCT (w)
 Mutants: ATTAACCACTCcCGGGAGCTCT (c)
 Mutants: ATTAACCACTCgCGGGAGCTCT (g)
 5 Mutants: ATTAACCACTCtCGGGAGCTCT (t)
 TGGTGNKYGCCCT (pooled probe).

The sixteen individual component probes of the pooled probe hybridize to the four possible target sequences as follows:

		TARGET			
		w	c	g	t
10	TGGTGAGcGCCCT	n	n	y	n
	TGGTGcGcGCCCT	n	n	n	n
	TGGTGgGcGCCCT	n	n	n	n
15	TGGTGtGcGCCCT	n	n	n	n
	TGGTGAtcGCCCT	n	n	n	n
	TGGTGctcGCCCT	n	n	n	n
	TGGTGgtcGCCCT	n	n	n	n
	TGGTGttcGCCCT	n	n	n	n
20	TGGTGAGTGCCCT	y	n	n	n
	TGGTGcGTGCCCT	n	n	n	n
	TGGTGgGTGCCCT	n	n	n	n
	TGGTGtGTGCCCT	n	n	n	n
	TGGTGAtTGCCCT	n	n	n	n
25	TGGTGctTGCCCT	n	n	n	n
	TGGTGgtTGCCCT	n	n	n	n
	TGGTGttTGCCCT	n	n	n	n

The pooled probe hybridizes according to the aggregate of its components:

Pool: TGGTGNKYGCCCT y n y n

Thus, as stated above, it can be seen that a pooled probe having a y at the interrogation position hybridizes to the wildtype target and one of the mutants. Similar tables can be drawn to illustrate the hybridization patterns of probe pools having other pooled nucleotides at the interrogation position.

The above strategy of using pooled probes to analyze a single base in a target sequence can readily be extended to analyze any number of bases. At this point, the purpose of including three pooled positions within each probe will become apparent. In the example that follows, ten pools of probes, each containing three pooled probe positions, can be used to analyze each of a contiguous sequence of eight nucleotides in a target sequence.

ATTAACCACTCACGGGAGCTCT Reference sequence
 ----- Readable nucleotides

Pools:

5 4 TAATTNKYGAGTG
 5 AATTGNKRAGTGC
 6 ATTGGNKRGTGCC
 7 TTGGTNMRTGCCC
 8 TGGTGNKYGCCCT
 10 9 GGTGANKRCCCTC
 10 GTGAGNKYCCTCG
 11 TGAGTNMYCTCGA
 12 GAGTGNMYTCGAG
 13 AGTGCNMYCGAGA
 15

In this example, the different pooled probes tile across the reference sequence, each pooled probe differing from the next by increments of one nucleotide. For each of the readable nucleotides in the reference sequence, there are three probe pools having a pooled interrogation position aligned with the readable nucleotide. For example, the 12th nucleotide from the left in the reference sequence is aligned with pooled interrogation positions in pooled probes 8, 9, and 10. Comparison of the hybridization intensities of these pooled probes reveals the identity of the nucleotide occupying position 12 in a target sequence.

	Targets	Pools		
		8	9	10
30 Wild:	ATTAACCACTCACGGGAGCTCT	Y	Y	Y
Mutants:	ATTAACCACTCcCGGGAGCTCT	N	Y	Y
Mutants:	ATTAACCACTCgCGGGAGCTCT	Y	N	Y
35 Mutants:	ATTAACCACTctCGGGAGCTCT	N	N	Y

Example Intensities:

	= lit cell	Wild				
	= blank cell	'C'				
40		'G'				
		'T'				
		None				

45 Thus, for example, if pools 8, 9 and 10 all light up, one knows the target sequence is wildtype, If pools, 9 and 10

light up, the target sequence has a C mutant at position 12. If pools 8 and 10 light up, the target sequence has a G mutant at position 12. If only pool 10 lights up, the target sequence has a t mutant at position 12.

5 The identity of other nucleotides in the target sequence is determined by a comparison of other sets of three pooled probes. For example, the identity of the 13th nucleotide in the target sequence is determined by comparing the hybridization patterns of the probe pools designated 9, 10 and
10 11. Similarly, the identity of the 14th nucleotide in the target sequence is determined by comparing the hybridization patterns of the probe pools designated 10, 11, and 12.

 In the above example, successive probes tile across the reference sequence in increments of one nucleotide, and each
15 probe has three interrogation positions occupying the same positions in each probe relative to the terminus of the probe (i.e., the 7, 8 and 9th positions relative to the 3' terminus). However, the trellis strategy does not require that probes tile in increments of one or that the
20 interrogation position positions occur in the same position in each probe. In a variant of trellis tiling referred to as "loop" tiling, a nucleotide of interest in a target sequence is read by comparison of pooled probes, which each have a pooled interrogation position corresponding to the nucleotide
25 of interest, but in which the spacing of the interrogation position in the probe differs from probe to probe. Analogously to the block tiling approach, this allows several nucleotides to be read from a target sequence from a collection of probes that are identical except at the
30 interrogation position. The identity in sequence of probes, particularly at their 3' termini, simplifies synthesis of the array and result in more uniform probe density per cell.

 To illustrate the loop strategy, consider a reference
35 sequence of which the 4, 5, 6, 7 and 8th nucleotides (from the 3' termini are to be read. All of the four possible nucleotides at each of these positions can be read from comparison of hybridization intensities of five pooled probes. Note that the pooled positions in the probes are different

(for example in probe 55, the pooled positions are 4, 5 and 6 and in probe 56, 5, 6 and 7).

	TAACCACTCACGGGAGCA	Reference sequence
55	ATTNKYGAGTGCC	
56	ATTGNKRAGTGCC	
57	ATTGGNKRGTGCC	
58	ATTRGTNMGTGCC	
59	ATTKRTGNGTGCC	

Each position of interest in the reference sequence is read by comparing hybridization intensities for the three probe pools that have an interrogation position aligned with the nucleotide of interest in the reference sequence. For example, to read the fourth nucleotide in the reference sequence, probes 55, 58 and 59 provide pools at the fourth position. Similarly, to read the fifth nucleotide in the reference sequence, probes 55, 56 and 59 provide pools at the fifth position. As in the previous trellis strategy, one of the three probes being compared has an N at the pooled position and the other two have M or K, and (2) R or Y and (3) W or S.

The hybridization pattern of the five pooled probes to target sequences representing each possible nucleotide substitution at five positions in the reference sequence is shown below. Each possible substitution results in a unique hybridization pattern at three pooled probes, and the identity of the nucleotide at that position can be deduced from the hybridization pattern.

			Pools				
Targets			55	56	57	58	59
5	Wild:	TAACCACTCACGGGAGCA	Y	Y	Y	Y	Y
	Mutant:	TAAgCACTCACGGGAGCA	Y	N	N	N	N
	Mutant:	TAAtCACTCACGGGAGCA	Y	N	N	Y	N
	Mutant:	TAAaCACTCACGGGAGCA	Y	N	N	N	Y
10	Mutant:	TAACgACTCACGGGAGCA	N	Y	N	N	N
	Mutant:	TAAcTACTCACGGGAGCA	N	Y	N	N	Y
	Mutant:	TAACaACTCACGGGAGCA	Y	Y	N	N	N
15	Mutant:	TAACCcCTCACGGGAGCA	N	Y	Y	N	N
	Mutant:	TAACCgCTCACGGGAGCA	Y	N	Y	N	N
	Mutant:	TAACcTCTCACGGGAGCA	N	N	Y	N	N
20	Mutant:	TAACCagTCACGGGAGCA	N	N	N	Y	N
	Mutant:	TAACCAtTCACGGGAGCA	N	Y	N	Y	N
	Mutant:	TAACCAaTCACGGGAGCA	N	N	Y	Y	N
25	Mutant:	TAACCACaCACGGGAGCA	N	N	N	N	Y
	Mutant:	TAACCACcCACGGGAGCA	N	N	Y	N	Y
	Mutant:	TAACCACgCACGGGAGCA	N	N	N	Y	Y

Many variations on the loop and trellis tilings can be created. All that is required is that each position in sequence must have a probe with a 'N', a probe containing one of R/Y, M/K or W/S, and a probe containing a different pool from that set, complementary to the wild type target at that position, and at least one probe with no pool at all at that position. This combination allows all mutations at that position to be uniquely detected and identified.

A further class of strategies involving pooled probes are termed coding strategies. These strategies assign code words from some set of numbers to variants of a reference sequence. Any number of variants can be coded. The variants can include multiple closely spaced substitutions, deletions or insertions. The designation letters or other symbols assigned to each variant may be any arbitrary set of numbers, in any order. For example, a binary code is often used, but codes to other bases are entirely feasible. The numbers are often assigned such that each variant has a designation having at least one digit and at least one nonzero value for that digit. For example, in a binary system, a variant assigned the number

101, has a designation of three digits, with one possible nonzero value for each digit.

The designation of the variants are coded into an array of pooled probes comprising a pooled probe for each nonzero value of each digit in the numbers assigned to the variants. For example, if the variants are assigned successive number in a numbering system of base m , and the highest number assigned to a variant has n digits, the array would have about $n \times (m-1)$ pooled probes. In general, $\log_m (3N+1)$ probes are required to analyze all variants of N locations in a reference sequence, each having three possible mutant substitutions. For example, 10 base pairs of sequence may be analyzed with only 5 pooled probes using a binary coding system. Each pooled probe has a segment exactly complementary to the reference sequence except that certain positions are pooled. The segment should be sufficiently long to allow specific hybridization of the pooled probe to the reference sequence relative to a mutated form of the reference sequence. As in other tiling strategies, segments lengths of 9-21 nucleotides are typical. Often the probe has no nucleotides other than the 9-21 nucleotide segment. The pooled positions comprise nucleotides that allow the pooled probe to hybridize to every variant assigned a particular nonzero value in a particular digit. Usually, the pooled positions further comprises a nucleotide that allows the pooled probe to hybridize to the reference sequence. Thus, a wildtype target (or reference sequence) is immediately recognizable from all the pooled probes being lit.

When a target is hybridized to the pools, only those pools comprising a component probe having a segment that is exactly complementary to the target light up. The identity of the target is then decoded from the pattern of hybridizing pools. Each pool that lights up is correlated with a particular value in a particular digit. Thus, the aggregate hybridization patterns of each lighting pool reveal the value of each digit in the code defining the identity of the target hybridized to the array.

As an example, consider a reference sequence having four positions, each of which can be occupied by three possible mutations. Thus, in total there are 4×3 possible variant forms of the reference sequence. Each variant is assigned a binary number binary numbers 0001-1100 and the wildtype reference sequence is assigned the binary number 1111.

		X	X	X	X	-	4
10	Positions						
	Target: TAAC	C=1111	A=1111	C=1111	T=1111		
	CACGGGAGCA						
		G=0001	C=0010	G=0011	A=0100		
		T=0101	G=0110	T=0111	C=1000		
		A=1001	T=1010	A=1011	G=1100		

15

A first pooled probe is designed by including probes that complement exactly each variant having a 1 in the first digit.

20

	target(1111):	TAAC	C	A	C	T	CACGGGAGCA
	Mutant(0001):	TAAC	g	A	C	T	CACGGGAGCA
	Mutant(0101):	TAAC	t	A	C	T	CACGGGAGCA
	Mutant(1001):	TAAC	a	A	C	T	CACGGGAGCA
25	Mutant(0011):	TAAC	C	A	g	T	CACGGGAGCA
	Mutant(0111):	TAAC	C	A	t	T	CACGGGAGCA
	Mutant(1101):	TAAC	C	A	a	T	CACGGGAGCA

30	First pooled probe						
	=	ATTG	[GCAT]	T	[GCAT]	A	GTGCCC
	=	ATTG	N	T	N	A	GTGCCC

Second, third and fourth pooled probes are then designed respectively including component probes that hybridize to each variant having a 1 in the second, third and fourth digit.

XXXX - 4 positions examined

40	Target:	TAACCACTCACGGGAGCA			
	Pool 1(1):	ATTGnTnAGTGCCC =	16 probes	(4x1x4x1)	
	Pool 2(2):	ATTGGnnAGTGCCC =	16 probes	(1x4x4x1)	
	Pool 3(4):	ATTGyrydGTGCCC =	24 probes	(2x2x2x3)	
	Pool 4(8):	ATTGmwmbGTGCCC =	24 probes	(2x2x2x3)	

The pooled probes hybridize to variant targets as follows:

Hybridization pattern:

		Pools			
	Targets	1	2	3	4
5	Wild(1111)	Y	Y	Y	Y
	Mutant(0001):	Y	N	N	N
	Mutant(0101):	Y	N	Y	N
	Mutant(1001):	Y	N	N	Y
10	Mutant(0010):	N	Y	N	N
	Mutant(0110):	N	Y	Y	N
	Mutant(1010):	N	Y	N	Y
15	Mutant(0011):	Y	Y	N	N
	Mutant(0111):	Y	Y	Y	N
	Mutant(1101):	Y	N	Y	Y
	Mutant(0100):	N	N	Y	N
20	Mutant(1000):	N	N	N	Y
	Mutant(1100):	N	N	Y	Y

The identity of a variant (i.e., mutant) target is read directly from the hybridization pattern of the pooled probes. For example the mutant assigned the number 0001 gives a hybridization pattern of NNNY with respect to probes 4, 3, 2 and 1 respectively.

In the above example, variants are assigned successive numbers in a numbering system. In other embodiments, sets of numbers can be chosen for their properties. If the codewords are chosen from an error-control code, the properties of that code carry over to sequence analysis. An error code is a numbering system in which some designations are assigned to variants and other designations serve to indicate errors that may have occurred in the hybridization process. For example, if all codewords have an odd number of nonzero digits ('binary coding+error detection'), any single error in hybridization will be detected by having an even number of pools lit.

40

Wild
Target: TAACCACTCACGGGAGCA

45 Pool 1(1): ATTGnAnAGTGCCC = 16 Probes (4x1x4x1)
 Pool 2(2): ATTGGnnAGTGCCC = 16 Probes (1X4X4X1)
 Pool 3(4): ATTGryrhGTGCCC = 24 Probes (2X2X2X3)
 Pool 4(8): ATTGkwkvGTGCCC = 24 Probes (2X2X2X3)

A fifth probe can be added to make the number of pools that hybridize to any single mutation odd.

Pool 5(c): ATTGdhsmGTGCCC = 36 probes (2x2x3x3)

5

Hybridization of pooled probes to targets

		Pool				
Target		1	2	3	4	5
10	Target(11111): TAACCACTCACGGGAGCA	Y	Y	Y	Y	Y
	Mutant(00001): TAACgACTCACGGGAGCA	Y	N	N	N	N
	Mutant(10101): TAACTACTCACGGGAGCA	Y	N	N	N	N
	Mutant(11001): TAACaACTCACGGGAGCA	Y	N	N	Y	Y
15	Mutant(00010): TAACCCcCTCACGGGAGCA	N	Y	N	N	N
	Mutant(10110): TAACCGCTCACGGGAGCA	N	Y	Y	N	Y
	Mutant(11010): TAACctCTCACGGGAGCA	N	Y	N	Y	Y
20	Mutant(10011): TAACCAgTCACGGGAGCA	Y	Y	N	N	Y
	Mutant(00111): TAACAtTCACGGGAGCA	Y	Y	Y	N	N
	Mutant(01101): TAACCAaTCACGGGAGCA	Y	N	Y	Y	N
	Mutant(00100): TAACCACaCACGGGAGCA	N	N	Y	N	N
	Mutant(01000): TAACCACcCCACGGGAGCA	N	N	N	Y	N
25	Mutant(11100): TAACCACgCACGGGAGCA	N	N	Y	Y	Y

9. Bridging Strategy

Probes that contain partial matches to two separate (i.e., non contiguous) subsequences of a target sequence sometimes hybridize strongly to the target sequence. In certain instances, such probes have generated stronger signals than probes of the same length which are perfect matches to the target sequence. It is believed (but not necessary to the invention) that this observation results from interactions of a single target sequence with two or more probes simultaneously. This invention exploits this observation to provide arrays of probes having at least first and second segments, which are respectively complementary to first and second subsequences of a reference sequence. Optionally, the probes may have a third or more complementary segments. These probes can be employed in any of the strategies noted above. The two segments of such a probe can be complementary to disjoint subsequences of the reference sequences or contiguous subsequences. If the latter, the two segments in the probe are inverted relative to the order of the complement of the

reference sequence. The two subsequences of the reference sequence each typically comprises about 3 to 30 contiguous nucleotides. The subsequences of the reference sequence are sometimes separated by 0, 1, 2 or 3 bases. Often the sequences, are adjacent and nonoverlapping.

For example, a wild-type probe is created by complementing two sections of a reference sequence (indicated by subscript and superscript) and reversing their order. The interrogation position is designated (*) and is apparent from comparison of the structure of the wildtype probe with the three mutant probes. The corresponding nucleotide in the reference sequence is the "a" in the superscripted segment.

Reference: 5' T_{GGCTA}^{CGAGG}AATCATCTGTTA

Probes: 3' GCTCC CCGAT (Probe from first probe set)
 3' GCACC CCGAT
 3' GCCCC CCGAT
 3' GCGCC CCGAT

The expected hybridizations are:

Match:

GCTCCCCGAT
 ... TGGCTACGAGGAATCATCTGTTA
GCTCCCCGAT

Mismatch:

GCTCCCCGAT
 ... TGGCTACGAGGAATCATCTGTTA
GCGCCCCGAT

Bridge tilings are specified using a notation which gives the length of the two constituent segments and the relative position of the interrogation position. The designation n/m indicates a segment complementary to a region of the reference sequence which extends for n bases and is located such that the interrogation position is in the mth base from the 5' end. If m is larger than n, this indicates that the entire segment is to the 5' side of the interrogation position. If m is negative, it indicates that the interrogation position is the absolute value of m bases 5' of the first base of the segment (m cannot be zero). Probes comprising multiple segments, such as n/m + a/b + ... have a first segment at the 3' end of the

probe and additional segments added 5' with respect to the first segment. For example, a 4/8 tiling consists of (from the 3' end of the probe) a 4 base complementary segment, starting 7 bases 5' of the interrogation position, followed by a 6 base region in which the interrogation position is located at the third base. Between these two segments, one base from the reference sequence is omitted. By this notation, the set shown above is a 5/3 + 5/8 tiling. Many different tilings are possible with this method, since the lengths of both segments can be varied, as well as their relative position (they may be in either order and there may be a gap between them) and their location relative to the interrogation position.

As an example, a 16 mer oligo target was hybridized to a chip containing all 4^{10} probes of length 10. The chip includes short tilings of both standard and bridging types. The data from a standard 10/5 tiling was compared to data from a 5/3 + 5/8 bridge tiling (see Table 1). Probe intensities (mean count/pixel) are displayed along with discrimination ratios (correct probe intensity / highest incorrect probe intensity). Missing intensity values are less than 50 counts. Note that for each base displayed the bridge tiling has a higher discrimination value.

TABLE 1: Comparison of Standard and Bridge Tilings

TILING	PROBE BASE:	CORRECT PROBE BASE			
		C	A	C	C
STANDARD (10/5)	A	92	496	294	299
	C	536	148	532	534
	G	69	167	72	52
	T	146	95	212	126
DISCRIMINATION:		3.7	3.0	1.8	1.8
BRIDGING 5/3 + 5/8	A	-	404	-	156
	C	276	-	345	379
	G	-	80	-	-
	T	-	-	-	58
DISCRIMINATION:		>5.5	5.1	2.4	1.26

45

The bridging strategy offers the following advantages:

(1) Higher discrimination between matched and mismatched probes,

(2) The possibility of using longer probes in a bridging tiling, thereby increasing the specificity of the hybridization, without sacrificing discrimination,

(3) The use of probes in which an interrogation position is located very off-center relative to the regions of target complementarity. This may be of particular advantage when, for example, when a probe centered about one region of the target gives low hybridization signal. The low signal is overcome by using a probe centered about an adjoining region giving a higher hybridization signal.

(4) Disruption of secondary structure that might result in annealing of certain probes (see previous discussion of helper mutations).

10. Deletion Tiling

Deletion tiling is related to both the bridging and helper mutant strategies described above. In the deletion strategy, comparisons are performed between probes sharing a common deletion but differing from each other at an interrogation position located outside the deletion. For example, a first probe comprises first and second segments, each exactly complementary to respective first and second subsequences of a reference sequence, wherein the first and second subsequences of the reference sequence are separated by a short distance (e.g., 1 or 2 nucleotides). The order of the first and second segments in the probe is usually the same as that of the complement to the first and second subsequences in the reference sequence. The interrogation position is usually separated from the comparison is performed with three other probes, which are identical to the first probe except at an interrogation position, which is different in each probe.

Reference: . . . AGTACCAGATCTCTAA . . .

Probe set: CATGGNC AGAGA (N = interrogation position).

Such tilings sometimes offer superior discrimination in hybridization intensities between the probe having an interrogation position complementary to the target and other probes. Thermodynamically, the difference between the hybridizations to matched and mismatched targets for the probe

set shown above is the difference between a single-base bulge, and a large asymmetric loop (e.g., two bases of target, one of probe). This often results in a larger difference in stability than the comparison of a perfectly matched probe with a probe showing a single base mismatch in the basic tiling strategy.

The superior discrimination offered by deletion tiling is illustrated by Table 2, which compares hybridization data from a standard 10/5 tiling with a (4/8 + 6/3) deletion tiling of the reference sequence. (The numerators indicate the length of the segments and the denominators, the spacing of the deletion from the far termini of the segments.) Probe intensities (mean count/pixel) are displayed along with discrimination ratios (correct probe intensity / highest incorrect probe intensity). Note that for each base displayed the deletion tiling has a higher discrimination value than either standard tiling shown.

TABLE 2. Comparison of Standard and Deletion Tilings

TILING	PROBE BASE:	CORRECT PROBE BASE			
		C	A	C	C
STANDARD (10/5)	A	92	496	294	299
	C	536	148	532	534
	G	69	167	72	52
	T	146	95	212	126
DISCRIMINATION:		3.7	3.0	1.8	1.8
DELETION 4/8 + 6/3	A	6	412	29	48
	C	297	32	465	160
	G	8	77	10	4
	T	8	26	31	5
DISCRIMINATION:		37.1	5.4	15	3.3
STANDARD (10/7)	A	347	533	228	277
	C	729	194	536	496
	G	232	231	102	89
	T	344	133	163	150
DISCRIMINATION:		2.1	2.3	2.3	1.8

The use of deletion or bridging probes is quite general. These probes can be used in any of the tiling strategies of the invention. As well as offering superior discrimination, the use of deletion or bridging strategies is advantageous for

certain probes to avoid self-hybridization (either within a probe or between two probes of the same sequence)

C. Preparation of Target Samples

5 The target polynucleotide, whose sequence is to be determined, is usually isolated from a tissue sample. If the target is genomic, the sample may be from any tissue (except exclusively red blood cells). For example, whole blood, peripheral blood lymphocytes or PBMC, skin, hair or semen are
10 convenient sources of clinical samples. These sources are also suitable if the target is RNA. Blood and other body fluids are also a convenient source for isolating viral nucleic acids. If the target is mRNA, the sample is obtained from a tissue in which the mRNA is expressed. If the
15 polynucleotide in the sample is RNA, it is usually reverse transcribed to DNA. DNA samples or cDNA resulting from reverse transcription are usually amplified, e.g., by PCR. Depending on the selection of primers and amplifying enzyme(s), the amplification product can be RNA or DNA.
20 Paired primers are selected to flank the borders of a target polynucleotide of interest. More than one target can be simultaneously amplified by multiplex PCR in which multiple paired primers are employed. The target can be labelled at one or more nucleotides during or after amplification. For
25 some target polynucleotides (depending on size of sample), e.g., episomal DNA, sufficient DNA is present in the tissue sample to dispense with the amplification step.

 When the target strand is prepared in single-stranded form as in preparation of target RNA, the sense of the strand
30 should of course be complementary to that of the probes on the chip. This is achieved by appropriate selection of primers. The target is preferably fragmented before application to the chip to reduce or eliminate the formation of secondary structures in the target. The average size of targets
35 segments following hybridization is usually larger than the size of probe on the chip.

II. ILLUSTRATIVE CHIPS

A. HIV Chip

HIV has infected a large and expanding number of people, resulting in massive health care expenditures. HIV can rapidly become resistant to drugs used to treat the infection, primarily due to the action of the heterodimeric protein (51 kDa and 66 kDa) HIV reverse transcriptase (RT) both subunits of which are encoded by the 1.7 kb pol gene. The high error rate (5-10 per round) of the RT protein is believed to account for the hypermutability of HIV. The nucleoside analogues, i.e., AZT, ddI, ddC, and d4T, commonly used to treat HIV infection are converted to nucleotide analogues by sequential phosphorylation in the cytoplasm of infected cells, where incorporation of the analogue into the viral DNA results in termination of viral replication, because the 5' -> 3' phosphodiester linkage cannot be completed. However, after about 6 months to 1 year of treatment or less, HIV typically mutates the RT gene so as to become incapable of incorporating the analogue and so resistant to treatment. Several mutations known to be associated with drug resistance are shown in the table below. After a virus having drug resistance via a mutation becomes predominant, the patient suffers dramatically increased viral load, worsening symptoms (typically more frequent and difficult-to-treat infections), and ultimately death. Switching to a different treatment regimen as soon as a resistant mutant virus takes hold may be an important step in patient management which prolongs patient life and reduces morbidity during life.

TABLE 3
SOME RT MUTATIONS ASSOCIATED WITH DRUG RESISTANCE

ANTIVIRAL	CODON	aa CHANGE	nt CHANGE
AZT	67	Asp -> Asn	GAC -> AAC
AZT	70	Lys -> Arg	AAA -> AGA
AZT	215	Thr -> Phe or Tyr	ACC -> TTC or TAC
AZT	219	Lys -> Gln or Glu	AAA -> CAA or GAA
AZT	41	Met -> Leu	ATG -> TTG or CTG
ddI and ddC	184	Met -> Val	ATG -> GTG
ddI and ddC	74	Leu -> Val	
TIBO 82150	100	Leu -> Ile	
ddC	65	Lys -> Asn	AAA -> AGA
ddC	69	Thr -> Asp	ACT -> GAT
3TC	184	Met -> Val	ATG -> GTG or GTA
3TC	184	Met -> Ile	ATG -> ATA
AZT + ddI	62	Ala -> Val	GCC -> GTC
AZT + ddI	75	Val -> Ile	GTA -> ATA
AZT + ddI	77	Phe -> Leu	TTC -> TTA
AZT + ddI	116	Phe -> Tyr	TTT -> TAT
AZT + ddI	151	Gln -> Met	CAG -> ATG
Nevaripine	103	Lys -> Asn	AAA -> AAT
	106	Val -> Ala	GTA -> GCA
	108		
	181	Tyr -> Cys	TAT -> TGT
	188	Tyr -> His	TAT -> CAT
	190	Gly -> Ala	GGA -> GCA

N.B.. Other mutations confer resistance to other drugs.

A second important therapeutic target for anti-HIV drugs is the aspartyl protease enzyme encoded by the HIV genome, whose function is required for the formation of infectious progeny. See Robbins & Plattner, *J. Acquired Immune Deficiency Syndromes* 6, 162-170 (1993); Kozal et al., *Curr. Op. Infect. Dis.* 7:72-81 (1994). The protease function in

processing of viral precursor polypeptides to their active forms. Drugs targeted against this enzyme do not impair endogenous human proteases, thereby achieving a high degree of selective toxicity. Moreover, the protease is expressed later in the life-cycle that reverse transcriptase, thereby offering the possibility of a combined attack on HIV at two different times in its life-cycle. As for drugs targeted against the reverse transcriptase, administration of drugs to the protease can result in acquisition of drug resistance through mutation of the protease. By monitoring the protease gene from patients, it is possible to detect the occurrence of mutations, and thereby make appropriate adjustments in the drug(s) being administered.

In addition to being infected with HIV, AIDS patients are often also infected with a wide variety of other infectious agents giving rise to a complex series of symptoms. Often diagnosis and treatment is difficult because many different pathogens (some life-threatening, others routine) cause similar symptoms. Some of these infections, so-called opportunistic infections, are caused by bacterial, fungal, protozoan or viral pathogens which are normally present in small quantity in the body, but are held in check by the immune system. When the immune system in AIDS patients fails, these normally latent pathogens can grow and generate rampant infection. In treating such patients, it would be desirable simultaneously to diagnose the presence or absence of a variety of the most lethal common infections, determine the most effective therapeutic regime against the HIV virus, and monitor the overall status of the patient's infection.

The present invention provides DNA chips for detecting the multiple mutations in HIV genes associated with resistance to different therapeutics. These DNA chips allow physicians to monitor mutations over time and to change therapeutics if resistance develops. Some chips also provide probes for diagnosis of pathogenic microorganisms that typically occur in AIDS patients.

The sequence selected as a reference sequence can be from anywhere in the HIV genome, but should preferably cover a

region of the HIV genome in which mutations associated with drug resistance are known to occur. A reference sequence is usually between about 5, 10, 20, 50, 100, 500, 1000, 5,000 or 10,000 bases in length, and preferably is about 100-1700 bases in length. Some reference sequences encompass at least part of the reverse transcriptase sequence encoded by the pol gene. Preferably, the reference sequence encompasses all, or substantially all (i.e, about 75 or 90%) of the reverse transcriptase gene. Reverse transcriptase is the target of several drugs and as noted, above, the coding sequence is the site of many mutations associated with drug resistance. In some chips, the reference sequence contains the entire region coding reverse transcriptase (850 bp), and in other chips, subfragments thereof. In some chips, the reference sequence includes other subfragments of the pol gene encoding HIV protease or endonuclease, instead of, or as well as the segment encoding reverse transcriptase. In some chips, the reference sequence also includes other HIV genes such as env or gag as well as or instead of the reverse transcriptase gene. Certain regions of the gag and env genes are relatively well conserved, and their detection provides a means for identifying and quantifying the amount of HIV virus infecting a patient. In some chips, the reference sequence comprises an entire HIV genome.

It is not critical from which strain of HIV the reference sequence is obtained. HIV strains are classified as HIV-I, HIV-II or HIV-III, and within these generic groupings there are several strains and polymorphic variants of each of these. BRU, SF2, HXB2, HXB2R are examples of HIV-1 strains, the sequences of which are available from GenBank. The reverse transcriptase genes of the BRU and SF2 strains differ at 23 nucleotides. The HXB2 and HXB2R strains have the same reverse transcriptase gene sequence, which differs from that of the BRU strain at four nucleotides, and that of SF2 by 27 nucleotides. In some chips, the reference sequence corresponds exactly to the reverse transcriptase sequence in the wildtype version of a strain. In other chips, the reference sequence corresponds to a consensus sequence of

several HIV strains. In some chips, the reference sequence corresponds to a mutant form of a HIV strain.

Chips are designed in accordance with the tiling strategies noted above. The probes are designed to be complementary to either the coding or noncoding strand of the HIV reference sequence. If only one strand is to be read, it is preferable to read the coding strand. The greater percentage of A residues in this strand relative to the noncoding strand generally result in fewer regions of ambiguous sequence.

Some chips contain additional probes or groups of probes designed to be complementary to a second reference sequence. The second reference sequence is often a subsequence of the first reference sequence bearing one or more commonly occurring HIV mutations or interstrain variations (e.g., within codons 67, 70, 215 or 219 of the reverse transcriptase gene). The inclusion of a second group is particularly useful for analyzing short subsequences of the primary reference sequence in which multiple mutations are expected to occur within a short distance commensurate with the length of the probes (i.e., two or more mutations within 9 to 21 bases).

The total number of probes on the chips depends on the tiling strategy, the length of the reference sequence and the options selected with respect to inclusion of multiple probe lengths and secondary groups of probes to provide confirmation of the existence of common mutations. To read much or all of the HIV reverse transcriptase gene (857 b for the BRU strain), chips tiled by the basic strategy typically contain at least $857 \times 4 = 3428$ probes.

The target HIV polynucleotide, whose sequence is to be determined, is usually isolated from blood samples (peripheral blood lymphocytes or PBMC) in the form of RNA. The RNA is reverse transcribed to DNA, and the DNA product is then amplified. Depending on the selection of primers and amplifying enzyme, the amplification product can be RNA or DNA. Suitable primers for amplification of target are shown in the table below.

TABLE 4
AMPLIFICATION OF TARGET

TARGET SIZE	FORWARD PRIMER	REVERSE PRIMER
1,742 bp	GTAGAATTCTGTTGACTCAGATTGG	GATAAGCTTGGGCCTTATCTATTCCAT
535 bp	AAATCCATACAATACTCCAGTATTTC	ACCCATCCAAAGGAATGGAGGTTCTTTC
323 bp	Genbank # K02013 1889-1908	bases 2211-2192
	AATTAACCCTCACTAAAGGGGaga ggaagaatctgttgactcagattggt (RT#1-T3)	AATTTAATACGACTCACTATAGGGGAttccc ctaacttctgtatgcatgaca-3' (89-391 T7)
	AATTAACCCTCACTAAAGGGGaga agtatactgcattaccatacctagta (RT#3-T3)	
	TAATACGACTCACTATAGGGGAGA tcgacgcaggactcggcttgctgaa (HV1-T2)	
	AATTAACCCTCACTAAAGGGGAGA ccttgtaagtcattggtcttaaggta (HV2-T3)	

In another aspect of the invention, chips are provided for simultaneous detection of HIV and microorganisms that commonly parasitize AIDS patients (e.g., cytomegalovirus (CMV), Pneumocystis carini (PCP), fungi (candida albicans), mycobacteria). Non-HIV viral pathogens are detected and their drug resistance determined using a similar strategy as for HIV. That is groups of probes are designed to show complementarity to a target sequence from a region of the genome of a nonviral pathogen known to be associated with acquisition of drug resistance. For example, CMV and HSV viruses, which frequently co-parasitize AIDS patients, undergo mutations to acquire resistance to acyclovir.

For detection of non-viral pathogens, the chips include an array of probes which allow full-sequence determination of 16S ribosomal RNA or corresponding genomic DNA of the pathogens. The additional probes are designed by the same principles as described above except that the target sequence is a variable region from a 16S RNA (or corresponding DNA) of a pathogenic microorganism. Alternatively, the target sequence can be a consensus sequences of variable 16S rRNA regions from multiple organisms. 16S ribosomal DNA and RNA is present in all organisms (except viruses) and the sequence of the DNA or RNA is closely related to the evolutionary genetic distance between any two species. Hence, organisms which are quite close in type (e.g., all mycobacteria) share a common

region of 16S rDNA, and differ in other regions (variable regions) of the 16S rRNA. These differences can be exploited to allow identification of the different subtype strains. The full sequence of 16S ribosomal RNA or DNA read from the chip is compared against a database of the sequence of thousands of known pathogens to type unambiguously most nonviral pathogens infecting AIDS patients.

In a further embodiment, the invention provides chips which also contain probes for detection of bacterial genes conferring antibiotic resistance. An antibiotic resistance gene can be detected by hybridization to a single probe employed in a reverse dot blot format. Alternatively, a group of probes can be designed according to the same principles discussed above to read all or part the DNA sequence encoding an antibiotic resistance gene. Analogous probes groups are designed for reading other antibiotic resistance gene sequences. Antibiotic resistance frequently resides in one of the following genes in microorganisms coparasitizing AIDS patients: *rpoB* (encoding RNA polymerase), *katG* (encoding catalase peroxidase, and DNA gyrase A and B genes.

The inclusion of probes for combinations of tests on a single chip simulates the clinical diagnosis tree that a physician would follow based on the presentation of a given syndrome which could be caused by any number of possible pathogens. Such chips allow identification of the presence and titer of HIV in a patient, identification of the HIV strain type and drug resistance, identification of opportunistic pathogens, and identification of the drug resistance of such pathogens. Thus, the physician is simultaneously apprised of the full spectrum of pathogens infecting the patient and the most effective treatments therefor.

Exemplary HIV Chips(a) HV 273

The HV 273 chip contains an array of oligonucleotide probes for analysis of an 857 base HIV amplicon between nucleotides 2090 and 2946 (HIVBRU strain numbering). The chip contains four groups of probes: 11 mers, 13 mers, 15 mers and 17 mers. From top to bottom, the HV 273 chip is occupied by rows of 11 mers, followed by rows of 13 mers, followed by rows of 15 mers followed by rows of 17 mers. The interrogation position is nucleotide 6, 7, 8 and 9 respectively in the different sized chips. This arrangement of the different sized probes is referred to as being "in series." Within each size group, there are four probe sets laid down in an A-lane, a C-lane a G-lane and a T-lane respectively. Each lane contains an overlapping series of probes with one probe for each nucleotide in the 2090-2946 HIV reverse transcriptase reference sequence. (i.e., 857 probes per lane). The lanes also include a few column positions which are empty or occupied by control probes. These positions serve to orient the chip, determine background fluorescence and punctuate different subsequences within the target. The chip has an area of 1.28 x 1.28 cm, within which the probes form a 130 X 135 matrix (17,550 cells total). The area occupied by each probe (i.e., a probe cell) is about 98 X 95 microns.

The chip was tested for its capacity to sequence a reverse transcriptase fragment from the HIV strain SF2. An 831 bp RNA fragment (designated pPol19) spanning most of the HIV reverse transcriptase coding sequence was amplified by PCR, using primers tagged with T3 and T7 promoter sequences. The primers, designated RT#1-T3 and 89-391 T7 are shown in Table 4; see also Gingeras et al., *J. Inf. Dis.* 164, 1066-1074 (1991) (incorporated by reference in its entirety for all purposes). RNA was labelled by incorporation of fluorescent nucleotides. The RNA was fragmented by heating and hybridized to the chip for 40 min at 30 degrees. Hybridization signals were quantified by fluorescence imaging.

Taking the best data from the four probes sets at each position in the target sequence, 715 out of 821 bases were

read correctly (87%). (Comparisons are based on the sequence of pPol19 determined by the conventional dideoxy method to be identical to SF2). In general, the longer sized probes yielded more sequence than the shorter probes. Of the 21 positions at which the SF2 and BRU strains diverged within the target, 19 were read correctly.

Many of the short ambiguous regions in the target arise in segments of the target flanking the points at which the SF2 and BRU sequences diverge. These ambiguities arise because in these regions the comparison of hybridization signals is not drawn between perfectly matched and single base mismatch probes but between a single-mismatched probe and three probes having two mismatches. These ambiguities in reading an SF2 sequence would not detract from the chip's ability to read a BRU sequence either alone or in a mixture with an SF2 target sequence.

In a variation of the above procedure, the chip was treated with RNase after hybridization of the pPol19 target to the probes. Addition of RNase digests mismatched target and thereby increases the signal to noise ratio. RNase treatment increased the number of correctly read bases to 743/821 or 90% (combining the data from the four groups of probes).

In a further variation, the RNA target was replaced with a DNA target containing the same segment of the HIV genome. The DNA probe was prepared by linear amplification using Taq polymerase, RT#1-T3 primer, and fluorescein d-UTP label. The DNA probe was fragmented with uracil DNA glycosylase and heat treatment. The hybridization pattern across the array and percentage of readable sequence were similar to those obtained using an RNA target. However, there were a few regions of sequence that could be read from the RNA target that could not be read from the DNA target and vice versa.

(b) HV 407 Chip

The 407 chip was designed according to the same principles as the HV 273 chip, but differs in several respects. First, the oligonucleotide probes on this chip are designed to exhibit perfect sequence identity (with the

exception of the interrogation position on each probe) to the HIV strain SF2 (rather than the BRU strain as was the case for the HV 273 chip). Second, the 407 chip contains 13 mers, 15 mers, 17 mers and 19 mers (with interrogation positions at nucleotide 7, 8, 9 and 10 respectively), rather than the 11 mers, 13 mers, 15 mers and 17 mers on the HV 273 chip. Third, the different sized groups of oligomers are arranged in parallel in place of the in-series arrangement on the HV 273 chip. In the parallel arrangement, the chip contains from top to bottom a row of 13 mers, a row of 15 mers, a row of 17 mers, a row of 19 mers, followed by a further row of 13 mers, a row of 15 mers, a row of 17 mers, a row of 19 mers, followed by a row of 13 mers, and so forth. Each row contains 4 lanes of probes, an A lane, a C lane, a G lane and a T lane, as described above. The probes in each lane tile across the reference sequence. The layout of probes on the HV 407 chip is shown in Fig. 10.

The 407 chip was separately tested for its ability to sequence two targets, pPol19 RNA and 4MUT18 RNA. pPol19 contains an 831 bp fragment from the SF2 reverse transcriptase gene which exhibits perfect complementarity to the probes on the 407 chip (except of course for the interrogation positions in three of the probes in each column). 4MUT18 differs from the reference sequence at thirty-one positions within the target, including five positions in codons 67, 70, 215 and 219 associated with acquisition of drug resistance. Target RNA was prepared, labelled and fragmented as described above and hybridized to the HV 407 chip. The hybridization pattern for the pPol19 target is shown in Fig. 11.

The sequences read off the chip for the pPol19 and 4MUT18 targets are both shown in Fig. 12 (although the two sequences were determined in different experiments). The sequence labelled wildtype in the Figure is the reference sequence. The four lanes of sequence immediately below the reference sequence are the respective sequences read from the four-sized groups of probes for the pPol19 target (from top-to-bottom, 13 mers, 15 mers, 17 mers and 19 mers). The next four lanes of sequence are the sequences read from the four-sized groups of

probes for the 4MUT18 target (from top-to-bottom in the same order). The regions of sequences shown in normal type are those that could be read unambiguously from the chip. Regions where sequence could not be accurately read are shown

5 highlighted. Some regions of sequence that could not be read from one sized set of probes could be read from another.

Taking the best result from the four sized groups of probes at each column position, about 97% of bases in the pPol19 sequence and about 90% of bases in the 4MUT18 sequence
10 were read accurately. Of the 31 nucleotide differences between 4MUT18 and the reference sequence, twenty-seven were read correctly including three of the nucleotide changes associated with acquisition of drug resistance. Of the
15 ambiguous regions in the 4MUT18 sequence determination, most occurred in the 4MUT18 segments flanking points of divergence between the 4MUT18 and reference sequences. Notably, most of the common mutations in HIV reverse transcriptase associated with drug resistance (see Table 3) occur at sequence positions that can be read from the chip. Thus, most of the commonly
20 occurring mutations can be detected by a chip containing an array of probes based on a single reference sequence.

Comparison of the sequence read of the probes of different sizes is useful in determining the optimum size probe to use for different regions of the target. The
25 strategy of customizing probe length within a single group of probe sets minimizes the total number of probes required to read a particular target sequence. This leaves ample capacity for the chip to include probes to other reference sequences (e.g., 16S RNA for pathogenic microorganisms) as discussed
30 below.

The HV 407 chip has also been tested for its capacity to detect mixtures of different HIV strains. The mixture comprises varying proportions of two target sequences; one a
segment of a reverse transcriptase gene from a wildtype SF2
35 strain, the other a corresponding segment from an SF2 strain bearing a codon 67 mutation. See Fig. 13. The Figure also represents the probes on the chip having an interrogation position for reading the nucleotide in which the mutation

occurs. A single probe in the Figure represents four probes on the chip with the symbol (o) indicating the interrogation position, which differs in each of the four probes. Figure 14 shows the fluorescence intensity for the four 13 mers and the four 15 mers having an interrogation position for reading the nucleotide in the target sequence in which the mutation occurs. As the percentage of mutant target is increase, the fluorescence intensity of the probe exhibiting perfect complementarity to the wildtype target decreases, and the intensity of the probe exhibiting perfect complementarity to the mutant sequence increases. The intensities of the other two probes do not change appreciably. It is concluded that the chip can be used to analyze simultaneously a mixture of strains, and that a strain comprising as little as ten percent of a mixture can be easily detected.

c. Protease Chip

A protease chip was constructed using the basic tiling strategy. The chip comprises four probes tiling across a 382 nucleotide span including 297 nucleotides from the protease coding sequence. The reference sequence was a consensus Clay-B HIV protease sequence. Different probes lengths were employed for tiling different regions of the reference sequence. Probe lengths were 11, 14, 17 and 20 nucleotides with interrogation positions at or adjacent to the center of each probe. Lengths were optimized from prior hybridization data employing a chip having multiple tilings, each with a different probe length.

The chip was hybridized to four different single-stranded DNA protease target sequences (HXB2, SF2, NY5, pPol4mut18). Both sense and antisense strands were sequenced. Data from the chip was compared with that from an ABI sequencer. The overall accuracy from sequencing the four targets is illustrated in the Table 5 below.

Table 5

		ABI		Protease Chip	
		Sense	Antisense	Sense	Antisense
5	No call	0	4	9	4
	Ambiguous	6	14	17	8
	Wrong call	2	3	3	1
	TOTAL	8	21	29	13

10

ABI (sense) - 99.5%
Chip (sense) - 98.1%

15

ABI (antisense) - 98.6%
Chip (antisense) - 99.1%

20 Combining the data from sense and antisense strands, both the chip and the ABI sequencer provided 100% accurate data for all of the sequence from all four clones.

In a further test, the chip was hybridized to protease target sequences from viral isolates obtained from four patients before and after ddI treatment. The sequence read from the chip is shown in Fig. 15. Several mutations (indicated by arrows) have arisen in the samples obtained posttreatment. Particularly noteworthy was the chip's capacity to read a g/a mutation at nucleotide 207, notwithstanding the presence of two additional mutations (gt) at adjacent positions.

30

B. Cystic Fibrosis Chips

A number of years ago, cystic fibrosis, the most common severe autosomal recessive disorder in humans, was shown to be associated with mutations in a gene thereafter named the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. The CFTR gene is about 250 kb in size and has 27 exons. Wildtype genomic sequence is available for all exonic regions and exons/intron boundaries (Zielenski et al., *Genomics* 10, 214-228 (1991). The full-length wildtype cDNA sequence has also been described (see Riordan et al., *Science* 245, 1059-1065 (1989). Over 400 mutations have been mapped (see Tsui et al., *Hu. Mutat.* 1, 197-203 (1992). Many of the more common mutations are shown in Table 6. The most common cystic

45

fibrosis mutation is a three-base deletion resulting in the omission of amino acid #508 from the CFTR protein. The frequency of mutations varies widely in populations of different geographic or ethnic origin (see column 4 of Table 6). About 90% of all mutations having phenotypic effects occur in coding regions.

Detection of CFTR mutations is useful in a number of respects. For example, screening of populations can identify asymptomatic heterozygous individuals. Such individuals are at risk of giving rise to affected offspring suffering from CF if they reproduce with other such individuals. *In utero* screening of fetuses is also useful in identifying fetuses bearing 2 CFTR mutations. Identification of such mutations offers the possibility of abortion, or gene therapy. For couples known to be at risk of giving rise to affected progeny, diagnosis can be combined with *in vitro* reproduction procedures to identify an embryo having at least one wildtype CF allele before implantation. Screening children shortly after birth is also of value in identifying those having 2 copies of the defective gene. Early detection allows administration of appropriate treatment (e.g., Pulmozyme Antibiotics, Pertussive Therapy) thereby improving the quality of life and perhaps prolonging the life expectancy of an individual.

The source of target DNA for detecting of CFTR mutations is usually genomic. In adults, samples can conveniently be obtained from blood or mouthwash epithelial cells. In fetuses, samples can be obtained by several conventional techniques such as amniocentesis, chorionic villus sampling or fetal blood sampling. At birth, blood from the amniotic chord is a useful tissue source.

The target DNA is usually amplified by PCR. Some appropriate pairs of primers for amplifying segments of DNA including the sites of known mutations are listed in Tables 5 and 6.

Table 7

5

10

15

20

25

OLIGO NUMBER	SEQUENCE
787	TCTCCTTGGATATACTTGTGTGAATCAA
788	TCACCAGATTTCGTAGTCTTTTCATA
851	GTCTTGTGTTGAAATTCTCAGGGTAT
769	CTTGTACCAGCTCACTACCTAAT
887	ACCTGAGAAGATAGTAAGCTAGATGAA
888	AACTCCGCCTTTCCAGTTGTAT
934	TTAGTTTCTAGGGGTGGAAGATACA
935	TTAATGACACTGAAGATCACTGTTCTAT
789	CCATTCCAAGATCCCTGATATTTGAA
790	GCACATTTTTGCAAAGTTCATTAGA
891	TCATGGGCCATGTGCTTTTCAA
892	ACCTTCCAGCACTACAACTAGAA
760	CAAGTGAATCCTGAGCGTGATTT
850	GGTAGTGTGAAGGGTTCATATGCATA
762	GATTACATTAGAAGGAAGATGTGCCTTT
763	ACATGAATGACATTTACAGCAAATGCTT
931	GTGACCATATTGTAATGCATGTAGTGA
932	ATGGTGAACATATTTCTCAAGAGGTAA
955	TGT CTC TGT AAA CTG ATG GCT AAC A
884	TCGTATAGAGTTGATTGGATTGAGAA
885	CCATTAACCTAATGTGGTCTCATCACAA
886	CTACCATAATGCTTGGGAGAAATGAA
782	TCAAAGAATGGCACCAGTGTGAAA
901	TGCTTAGCTAAAGTTAATGAGTTCAT

OLIGO NUMBER	SEQUENCE
784	AATTGTGAAATTGTCTGCCATTCTTAA
785	GATTCACCTTACTGAACACAGTCTAACAA
791	AGGCTTCTCAGTGATCTGTTG
792	GAATCATTTCAGTGGGTATAAGCA
1013	GCCATGGTACCTATATGTCACAGAA
1012	TGCAGAGTAATATGAATTTCTTGAGTACA
766	GGGACTCCAAATATTGCTGTAGTAT
1065	GTACCTGTTGCTCCAGGTATGTT

Other primers can be readily devised from the known genomic and cDNA sequences of CFTR. The selection of primers, of course, depends on the areas of the target sequence that are to be screened. The choice of primers also depends on the strand to be amplified. For some regions of the CFTR gene, it makes little difference to the hybridization signal whether the coding or noncoding strand is used. In other regions, one strand may give better discrimination in hybridization signals between matched and mismatched probes than the other. The upper limit in the length of a segment that can be amplified from one pair of PCR primers is about 50 kb. Thus, for analysis of mutants through all or much of the CFTR gene, it is often desirable to amplify several segments from several paired primers. The different segments may be amplified sequentially or simultaneously by multiplex PCR. Frequently, fifteen or more segments of the CFTR gene are simultaneously amplified by PCR. The primers and amplification conditions are preferably selected to generate DNA targets. An asymmetric labelling strategy incorporating fluorescently labelled dNTPs for random labelling and dUTP for target fragmentation to an average length of less than 60 bases is preferred. The use of dUTP and fragmentation with

uracil N-glycosylase has the added advantage of eliminating carry over between samples.

Mutations in the CFTR gene can be detected by any of the tiling strategies noted above. The block tiling strategy is one particularly useful approach. In this strategy, a group (or block) of probes is used to analyze a short segment of contiguous nucleotides (e.g., 3, 5, 7 or 9) from a CFTR gene centered around the site of a mutation. The probes in a group are sometimes referred to as constituting a block because all probes in the group are usually identical except at their interrogation positions. As noted above, the probes may also differ in the presence of leading or trailing sequences flanking regions of complementarity. However, for ease of illustration, it will be assumed that such sequences are not present. As an example, to analyze a segment of five contiguous nucleotides from the CFTR gene, including the site of a mutation (such as one of the mutations in Table 6), a block of probes usually contains at least one wildtype probe and five sets of mutant probes, each having three probes. The wildtype probe has five interrogation positions corresponding to the five nucleotides being analyzed from the reference sequence. However, the identity of the interrogation positions is only apparent when the structure of the wildtype probe is compared with that of the probes in the five mutant probe sets. The first mutant probe set comprises three probes, each being identical to the wildtype probe, except in the first interrogation position, which differs in each of the three mutant probes and the wildtype probe. The second through fifth mutant probe sets are similarly composed except that the differences from the wildtype probe occur in the second through fifth interrogation position respectively. Note that in practice, each set of mutant probes is sometimes laid down on the chip juxtaposed with an associated wildtype probe. In this situation, a block would comprise five wildtype probes, each effectively providing the same information. However, visual inspection and confidence analysis of the chip is facilitated by the largely redundant information provided by five wildtype probes.

After hybridization to labelled target, the relative hybridization signals are read from the probes. Comparison of the intensities of the three probes in the first mutant probe set with that of the wildtype probe indicates the identity of the nucleotide in the target sequence corresponding to the first interrogation position. Comparison of the intensities of the three probes in the second mutant probe set with that of the wildtype probe indicates the identity of the nucleotide in the target sequence corresponding to the second interrogation position, and so forth. Collectively, the relative hybridization intensities indicate the identity of each of the five contiguous nucleotides in the reference sequence.

In a preferred embodiment, a first group (or block) of probes is tiled based on a wildtype reference sequence and a second group is tiled based a mutant version of the wildtype reference sequence. The mutation can be a point mutation, insertion or deletion or any combination of these. The combination of first and second groups of probes facilitates analysis when multiple target sequences are simultaneously applied to the chip, as is the case when a patient being diagnosed is heterozygous for the CFTR allele.

The above strategy is illustrated in Fig. 16, which shows two groups of probes tiled for a wildtype reference sequence and a point mutation thereof. The five mutant probe sets for the wildtype reference sequence are designated wt1-5, and the five mutant probe sets for the mutant reference sequence are designated m1-5. The letter N indicates the interrogation position, which shifts by one position in successive probe sets from the same group. The figure illustrates the hybridization pattern obtained when the chip is hybridized with a homozygous wildtype target sequence comprising nucleotides $n-2$ to $n+2$, where n is the site of a mutation. For the group of probes tiled based on the reference sequence, four probes are compared at each interrogation position. At each position, one of the four probes exhibits a perfect match with the target, and the other three exhibit a single-base mismatch. For the group of probes tiled based on the mutant

reference sequence, again four probes are compared at each interrogation position. At position, n, one probe exhibits a perfect match, and three probes exhibit a single base mismatch. Hybridization to a homozygous mutant yields an analogous pattern, except that the respective hybridization patterns of probes tiled on the wildtype and mutant reference sequences are reversed.

The hybridization pattern is very different when the chip is hybridized with a sample from a patient who is heterozygous for the mutant allele (see Fig. 17). For the group of probes tiled based on the wildtype sequence, at all positions but n, one probe exhibits a perfect match at each interrogation position, and the other three probes exhibit a one base mismatch. At position n, two probes exhibit a perfect match (one for each allele), and the other probes exhibit single-base mismatches. For the group of probes tiled on the mutant sequence, the same result is obtained. Thus, the heterozygote point mutant is easily distinguished from both the homozygous wildtype and mutant forms by the identity of hybridization patterns from the two groups of probes.

Typically, a chip comprises several paired groups of probes, each pair for detecting a particular mutation. For example, some chips contain 5, 10, 20, 40 or 100 paired groups of probes for detecting the corresponding numbers of mutations. Some chips are customized to include paired groups of probes for detecting all mutations common in particular populations (see Table 6). Chips usually also contain control probes for verifying that correct amplification has occurred and that the target is properly labelled.

The goal of the tiling strategy described above is to focus on short regions of the CFTR region flanking the sites of known mutation. Other tiling strategies analyze much larger regions of the CFTR gene, and are appropriate for locating and identifying hitherto uncharacterized mutations. For example, the entire genomic CFTR gene (250 kb) can be tiled by the basic tiling strategy from an array of about one million probes. Synthesis and scanning of such an array of probes is entirely feasible. Other tiling strategies, such as

the block tiling, multiplex tiling or pooling can cover the entire gene with fewer probes. Some tiling strategies analyze some or all of components of the CFTR gene, such as the cDNA coding sequence or individual exons. Analysis of exons 10 and 11 is particularly informative because these are location of many common mutations including the $\Delta F508$ mutation.

Exemplary CFTR chips

One illustrative chip bears an array of 1296 probes covering the full length of exon 10 of the CFTR gene arranged in a 36 x 36 array of 356 μm elements. The probes in the array can have any length, preferably in the range of from 10 to 18 residues and can be used to detect and sequence any single-base substitution and any deletion within the 192-base exon, including the three-base deletion known as $\Delta F508$. As described in detail below, hybridization of nanomolar concentrations of wild-type and $\Delta F508$ oligonucleotide target nucleic acids labeled with fluorescein to these arrays produces highly specific signals (detected with confocal scanning fluorescence microscopy) that permit discrimination between mutant and wild-type target sequences in both homozygous and heterozygous cases.

Sets of probes of a selected length in the range of from 10 to 18 bases and complementary to subsequences of the known wild-type CFTR sequence are synthesized starting at a position a few bases into the intron on the 5'-side of exon 10 and ending a few bases into the intron on the 3'-side. There is a probe for each possible subsequence of the given segment of the gene, and the probes are organized into a "lane" in such a way that traversing the lane from the upper left-hand corner of the chip to the lower righthand corner corresponded to traversing the gene segment base-by-base from the 5'-end. The lane containing that set of probes is, as noted above, called the "wild-type lane."

Relative to the wild-type lane, a "substitution" lane, called the "A-lane", was synthesized on the chip. The A-lane probes were identical in sequence to an adjacent (immediately below the corresponding) wild-type probe but contained, regardless of the sequence of the wild-type probe, a dA

residue at position 7 (counting from the 3'-end). In similar fashion, substitution lanes with replacement bases dC, dG, and dT were placed onto the chip in a "C-lane," a "G-lane," and a "T-lane," respectively. A sixth lane on the chip consisted of probes identical to those in the wild-type lane but for the deletion of the base in position 7 and restoration of the original probe length by addition to the 5'-end the base complementary to the gene at that position.

The four substitution lanes enable one to deduce the sequence of a target exon 10 nucleic acid from the relative intensities with which the target hybridizes to the probes in the various lanes. Various versions of such exon 10 DNA chips were made as described above with probes 15 bases long, as well as chips with probes 10, 14, and 18 bases long. For the results described below, the probes were 15 bases long, and the position of substitution was 7 from the 3'-end.

The sequences of several important probes are shown below. In each case, the letter "X" stands for the interrogation position in a given column set, so each of the sequences actually represents four probes, with A, C, G, and T, respectively, taking the place of the "X." Sets of shorter probes derived from the sets shown below by removing up to five bases from the 5'-end of each probe and sets of longer probes made from this set by adding up to three bases from the exon 10 sequence to the 5'-end of each probe, are also useful and provided by the invention.

```

3'-TTTATAXTAGAAACC
3'- TTATAGXAGAAACCA
3'- TATAGTXGAAACCAC
30 3'- ATAGTAXAAACCACA
3'- TAGTAGXAACCACAA
3'- AGTAGAXACCACAAA
3'- GTAGAAXCCACAAAG
3'- TAGAAAXCACAAAGG
35 3'- AGAAACXACAAAGGA
  
```

To demonstrate the ability of the chip to distinguish the Δ F508 mutation from the wild-type, two synthetic target

nucleic acids were made. The first, a 39-mer complementary to a subsequence of exon 10 of the CFTR gene having the three bases involved in the Δ F508 mutation near its center, is called the "wild-type" or wt508 target, corresponds to

5 positions 111-149 of the exon, and has the sequence shown below:

5'-CATTAAAGAAAATATCATCTTTGGTGTTCCTATGATGA.

The second, a 36-mer probe derived from the wild-type target by removing those same three bases, is called the "mutant"

10 target or mu508 target and has the sequence shown below, first with dashes to indicate the deleted bases, and then without dashes but with one base underlined (to indicate the base detected by the T-lane probe, as discussed below):

5'-CATTAAAGAAAATATCAT---TGGTGTTCCTATGATGA;

15 5'-CATTAAAGAAAATATCATTTGGTGTTCCTATGATGA.

Both targets were labeled with fluorescein at the 5'-end.

In three separate experiments, the wild-type target, the mutant target, and an equimolar mixture of both targets was exposed (0.1 nM wt508, 0.1 nM mu508, and 0.1 nM wt508 plus 0.1
20 nM mu508, respectively, in a solution compatible with nucleic acid hybridization) to a CF chip. The hybridization mixture was incubated overnight at room temperature, and then the chip was scanned on a reader (a confocal fluorescence microscope in photon-counting mode); images of the chip were constructed
25 from the photon counts) at several successively higher temperatures while still in contact with the target solution. After each temperature change, the chip was allowed to equilibrate for approximately one-half hour before being scanned. After each set of scans, the chip was exposed to
30 denaturing solvent and conditions to wash, i.e., remove target that had bound, the chip so that the next experiment could be done with a clean chip.

The results of the experiments are shown in Figures 18, 19, 20, and 21. Figure 18, in panels A, B, and C, shows an
35 image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to a wild-type target; in panel C, the chip was hybridized to a mutant Δ F508 target; and in panel B, the chip was hybridized to a mixture

of the wild-type and mutant targets. Figure 19, in sheets 1 - 3, corresponding to panels A, B, and C of Figure 3, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

These figures show that, for the wild-type target and the equimolar mixture of targets, the substitution probe with a nucleotide sequence identical to the corresponding wild-type probe bound the most target, allowing for an unambiguous assignment of target sequence as shown by letters near the points on the curve. The target wt508 thus hybridized to the probes in the wild-type lane of the chip, although the strength of the hybridization varied from probe-to-probe, probably due to differences in melting temperature. The sequence of most of the target can thus be read directly from the chip, by inference from the pattern of hybridization in the lanes of substitution probes (if the target hybridizes most intensely to the probe in the A-lane, then one infers that the target has a T in the position of substitution, and so on).

For the mutant target, the sequence could similarly be called on the 3'-side of the deletion. However, the intensity of binding declined precipitously as the point of substitution approached the site of the deletion from the 3'-end of the target, so that the binding intensity on the wild-type probe whose point of substitution corresponds to the T at the 3'-end of the deletion was very close to background. Following that pattern, the wild-type probe whose point of substitution corresponds to the middle base (also a T) of the deletion bound still less target. However, the probe in the T-lane of that column set bound the target very well. Examination of

the sequences of the two targets reveals that the deletion places an A at that position when the sequences are aligned at their 3'-ends and that the T-lane probe is complementary to the mutant target with but two mismatches near an end (shown below in lower-case letters, with the position of substitution underlined):

Target: 5'-CATTAAGAAAATATCATTGGTGTTCCTATGATGA

Probe: 3'-TagTAGTAACCCACAA

Thus the T-lane probe in that column set calls the correct base from the mutant sequence. Note that, in the graph for the equimolar mixture of the two targets, that T-lane probe binds almost as much target as does the A-lane probe in the same column set, whereas in the other column sets, the probes that do not have wild-type sequence do not bind target at all as well. Thus, that one column set, and in particular the T-lane probe within that set, detects the Δ F508 mutation under conditions that simulate the homozygous case and also conditions that simulate the heterozygous case.

Although in this example the sequence could not be reliably deduced near the ends of the target, where there is not enough overlap between target and probe to allow effective hybridization, and around the center of the target, where hybridization was weak for some other reason, perhaps high AT-content, the results show the method and the probes of the invention can be used to detect the mutation of interest. The mutant target gave a pattern of hybridization that was very similar to that of the wt508 target at the ends, where the two share a common sequence, and very different in the middle, where the deletion is located. As one scans the image from right to left, the intensity of hybridization of the target to the probes in the wild-type lane drops off much more rapidly near the center of the image for mu508 than for wt508; in addition, there is one probe in the T-lane that hybridizes intensely with mu508 and hardly at all with wt508. The results from the equimolar mixture of the two targets, which represents the case one would encounter in testing a heterozygous individual for the mutation, are a blend of the results for the separate targets, showing the power of the

invention to distinguish a wild-type target sequence from one containing the Δ F508 mutation and to detect a mixture of the two sequences.

The results above clearly demonstrate how the DNA chips
5 of the invention can be used to detect a deletion mutation, Δ F508; another model system was used to show that the chips can also be used to detect a point mutation as well. One mutation in the CFTR gene is G480C, which involves the replacement of the G in position 46 of exon 10 by a T,
10 resulting in the substitution of a cysteine for the glycine normally in position #480 of the CFTR protein. The model target sequences included the 21-mer probe wt480 to represent the wild-type sequence at positions 37-55 of exon 10:
5'-CCTTCAGAGGGTAAAATTAAG and the 21-mer probe mu480 to
15 represent the mutant sequence:
5'-CCTTCAGAGTGTAATAAATTAAG.

In separate experiments, a DNA chip was hybridized to each of the targets wt480 and mu480, respectively, and then scanned with a confocal microscope. Figure 20, in panels A,
20 B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to the wt480 target; in panel C, the chip was hybridized to the mu480 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets.
25 Figure 21, in sheets 1 - 3, corresponding to panels A, B, and C of Figure 20, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the
30 intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the
35 substitution probes ("2nd Highest").

These figures show that the chip could be used to sequence a 16-base stretch from the center of the target wt480 and that discrimination against mismatches is quite good

throughout the sequenced region. When the DNA chip was exposed to the target mu480, only one probe in the portion of the chip shown bound the target well: the probe in the set of probes devoted to identifying the base at position 46 in exon 10 and that has an A in the position of substitution and so is fully complementary to the central portion of the mutant target. All other probes in that region of the chip have at least one mismatch with the mutant target and therefore bind much less of it. In spite of that fact, the sequence of mu480 for several positions to both sides of the mutation can be read from the chip, albeit with much-reduced intensities from those observed with the wild-type target.

The results also show that, when the two targets were mixed together and exposed to the chip, the hybridization pattern observed was a combination of the other two patterns. The wild-type sequence could easily be read from the chip, but the probe that bound the mu480 target so well when only the mu480 target was present also bound it well when both the mutant and wild-type targets were present in a mixture, making the hybridization pattern easily distinguishable from that of the wild-type target alone. These results again show the power of the DNA chips of the invention to detect point mutations in both homo- and heterozygous individuals.

To demonstrate clinical application of the DNA chips of the invention, the chips were used to study and detect mutations in nucleic acids from genomic samples. Genomic samples from a individual carrying only the wild-type gene and an individual heterozygous for $\Delta F508$ were amplified by PCR using exon 10 primers containing the promoter for T7 RNA polymerase. Illustrative primers of the invention are shown below.

Exon Name Sequence

10	CFi9-T7	TAATACGACTCACTATAGGGAGatgacctaataatgatggggtt
10	CFi10c-T7	TAATACGACTCACTATAGGGAGtagtgtgaagggttcatatgc
35 10	CFi10c-T3	CTCGGAATTAACCCTCACTAAAGGtagtgtgaagggttcatatgc
11	CFi10-T7	TAATACGACTCACTATAGGGAGagcataactaaaagtgactctc
11	CFi11c-T7	TAATACGACTCACTATAGGGAGacatgaatgacatttacagcaa
11	CFi11c-T3	CGGAATTAACCCTCACTAAAGGacatgaatgacatttacagcaa

These primers can be used to amplify exon 10 or exon 11 sequences; in another embodiment, multiplex PCR is employed, using two or more pairs of primers to amplify more than one exon at a time.

5 The product of amplification was then used as a template for the RNA polymerase, with fluoresceinated UTP present to label the RNA product. After sufficient RNA was made, it was fragmented and applied to an exon 10 DNA chip for 15 minutes, after which the chip was washed with hybridization buffer and
10 scanned with the fluorescence microscope. A useful positive control included on many CF exon 10 chips is the 8-mer 3'-CGCCGCCG-5'. Figure 22, in panels A and B, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to nucleic acid
15 derived from the genomic DNA of an individual with wild-type $\Delta F508$ sequences; in panel B, the target nucleic acid originated from a heterozygous (with respect to the $\Delta F508$ mutation) individual. Figure 23, in sheets 1 and 2, corresponding to panels A and B of Figure 22, shows graphs of
20 fluorescence intensity versus tiling position.

These figures show that the sequence of the wild-type RNA can be called for most of the bases near the mutation. In the case of the $\Delta F508$ heterozygous carrier, one particular probe, the same one that distinguished so clearly between the
25 wild-type and mutant oligonucleotide targets in the model system described above, in the T-lane binds a large amount of RNA, while the same probe binds little RNA from the wild-type individual. These results show that the DNA chips of the invention are capable of detecting the $\Delta F508$ mutation in a
30 heterozygous carrier.

Further chips were constructed using the block tiling strategy to provide an array of probes for analyzing a CFTR mutation. The array comprised 93 mm x 96 μ m features arranged into eleven columns and four rows (44 total probes). Probes
35 in five of these columns were from four probe sets tiled based on the wildtype CFTR sequence and having interrogation positions corresponding to the site of a mutation and two bases on either side. Five of the remaining columns contained

four sets of probes tiled based on the mutant version of the CFTR sequence! These probe sets also had interrogation positions corresponding to the site of mutation and two nucleotides on either side. The eleventh column contained

5 four cells for control probes.

Fluorescently labeled hybridization targets were prepared by PCR amplification. 100 μ g of genomic DNA, 0.4 μ M of each primer, 50 μ M each dATP, dCTP, dGTP and dUTP (Pharmacia) in 10mM Tris-Cl, pH 8.3, 50 mM KCl, 2.5 mM MgCl₂ and 2 U Taq polymerase (Perkin-Elmer) were cycled 36 times using a Perkin-Elmer 9600 thermocycler and the following times and temperatures: 95°C, 10 sec., 55°C, 10 sec., 72°C, 30 sec. 10 μ l of this reaction product was used as a template in a second, asymmetric PCR reaction. Conditions included 1 μ M asymmetric PCR primer, 50 μ M each dATP, dCTP, TTP, 25 μ M fluorescein-dGTP (DuPont), 10 mM Tris-Cl, pH 9.1, 75 mM KCl, 3.5 mM MgCl₂. The reaction was cycled 5X with the following conditions: 95°C, 10 sec, 60°C, 10 sec, 55°C, 1 min. and 72°C, 1.5 min. This was immediately followed with another 20 cycles using the following conditions: 95°C, 10 sec, 60°C, 10 sec., 72°C, 1.5 min.

Amplification products were fragmented by treating with 2 U of Uracil-N-glycosylase (Gibco) at 30°C for 30 min. followed by heat denaturation at 95°C for 5 min. Finally, the labeled, fragmented PCR product was diluted into hybridization buffer made up of 5 X SSPE and 1 mM Cetyltrimethylammonium Bromide (CTAB). The dilution factor ranged from 10x to 25x with 40 μ l of sample being diluted into 0.4 ml to 1 ml of hybridization solution.

30 Target hybridization was generally carried out with the chip shaking in a small dish containing 500 μ l to 1 ml total volume of hybridization solution. All hybridizations were done at 30°C constant temperature. Alternatively, some hybridizations were carried out with chips enclosed in a plastic package with the 1 cm x 1 cm chip glued facing a 250 μ l fluid chamber. 250-350 μ l of hybridization solution was introduced and mixed using a syringe pump. Temperature was controlled by interfacing the back surface of the package with

a Peltier heating/cooling device. Following hybridization chips were washed with 5X SSPE, 0.1% Triton X-100 at 25°C-30°C prior to fluorescent image generation.

Hybridized, washed DNA chips were scanned for
5 fluorescence using a stage-scanning confocal epifluorescent
microscope and 488nm argon ion laser excitation. Emitted
light was collected through a band pass filter centered at
530nm. The resulting fluorescence image was spatially
reconstructed and intensity data were then analyzed. Features
10 with the peak fluorescence intensity in each column were
identified and compared with any signal intensity at the
remaining single base mismatch probe sites in the same column.
The sequences of the highest intensity features were then
compared across all ten columns of each sub-array to determine
15 whether peak intensity scores for the wild type sequence and
the mutant sequence were similar or significantly different.
These results were used to generate the genotype call of wild
type (high intensity signals only in wild type probe columns),
mutant (high intensity signals only in the mutant probe
20 columns) or heterozygous (high intensity signals in both the
wild type and mutant probe columns).

Figure 24 (panel A) shows an image of the fluorescence
signals in arrays designed to detect the G551D(G>A) and
Q552X(C>T) CFTR mutations. The hybridization target is an
25 exon 11 amplicon generated from wild type genomic DNA. Wild
type hybridization patterns are evident at both locations. No
significant fluorescence signal resulted at any of the
features with probes complementary to mutant or mismatched
sequences. Relative fluorescence intensities were six fold
30 brighter for the perfect matched wildtype features compared
with the background signal intensity at mutant and mismatch
features. In addition, the sequence at these loci can be
confirmed as AGGTC and GTCAA, respectively, where the bold
type face indicates the mutation sites. Figure 24 (panel B)
35 shows the same probe array features after hybridization with a
fluorescent target generated from DNA heterozygous for the
G551D mutation. Both the wild type and mutant probe columns
have features with significant fluorescence intensity,

indicating the hybridization of both wild type and mutant CFTR alleles at this site. Only wildtype probes hybridized with any significant fluorescence signal in the Q552X subarray indicating a wild type target sequence. However, an additional feature that did not hybridize in the first experiment shows significant fluorescence intensity in this experiment. Because the G551D and Q552X mutations are only two bases apart, the a probe sequence in the additional feature has a perfectly matched 12-mer overlap with the mutant G551D target.

Figure 25 (panels A and B) illustrates mutation analysis for $\Delta F508$, a three base pair deletion in Exon 10 of the CFTR gene. In contrast to the hybridization pattern seen in base change mutations, in mutations where bases are inserted or deleted, probe arrays show a different hybridization pattern. Identical probes are synthesized in the two central columns of base substitution arrays. As a result, either mutant or wild type target hybridizations always result in two side-by-side features (a doublet) with high fluorescence intensity at the center of the array. In a heterozygote hybridization, two sets of doublets, one matched to the wild type sequence and one to the mutant sequence occur (Figure 24, panel B). In contrast, wild type and mutant probe column sequences are offset from each other for deletion or insertion mutations and hybridization doublets are not seen. Instead of the six high intensity signals with one doublet, five independent features in alternating columns characterize a homozygote and ten features, one in each column will be positive with heterozygote targets. This is evident from the $\Delta F508$ hybridization pattern in Figure 25, panel A. Although a wildtype target has been hybridized and the highest intensity features confirm the wild type sequence (ATCTT), there is an additional hybridization in the first mutant column. Analysis of that probe sequence shows a 10 base perfect match with the mutant sequence.

The image in Figure 25, panel B resulted from hybridizing a DNA chip with a target homozygous for $\Delta F508$. In this image five features, all with probe sequences

complementary to the mutant show significant signal. The mutation sequence bridging the deletion site, ATTGG, is confirmed. Similar to what was seen in the example of the G551D mutation, there is added information in neighboring subarrays designed to detect the Δ I507 and F508C mutations. This is expected since they are in such close proximity to Δ F508 that their probe sets significantly overlap the Δ F508 probes. The Δ F508 homozygous target has no perfect matches with wild type or mutant probes in the Δ I507 and F508C subarrays. However, there are some low intensity signals within these two blocks of probes. The F508C array has a doublet that matches 11 bases of the mutant Δ F508 target. Similarly, the hybridization in the eighth column of the Δ I507 array has a probe that matches 13/14 bases with the target.

Figure 26 shows hybridization of a heterozygous double mutant Δ F508/F508C to the same array as described above. Conventional reverse dot blot would score this sample as a homozygous Δ F508 mutant. In the present assays, the Δ F508 and F508C alleles are separately detected by the respective subarrays designed to detect these mutations.

C. Chips for Cancer Diagnosis

There are at least two types of genes which are often altered in cancerous cells. The first type of gene is an oncogene such as a mismatch-repair gene, and the second type of gene is a tumor suppressor gene such as a transcription factor. Examples of mismatch repair oncogenes include hMSH2 (Fishel et al., *Cell* 75, 1027-1038 (1993)) and hMLH1 (Papadopoulos et al., *Science* 263, 1625-1628 (1994)). The most well-known example of a tumor suppressor gene is the p53 protein gene (Buchman et al., *Gene* 70, 245-252 (1988)). By monitoring the state of both oncogenes and tumor suppressor genes (individually and in combination) in a patient, it is possible to determine individual susceptibility to a cancer, a patient's prognosis upon cancer diagnosis, and to target therapy more efficiently.

The p53 gene spans 20 kbp in humans and has 11 exons, 10 of which are protein coding (see Tominaga et al., 1992,

Critical Reviews in Oncogenesis 3:257-282, incorporated herein by reference). The gene produces a 53 kilodalton phosphoprotein that regulates DNA replication. The protein acts to halt replication at the G1/S boundary in the cell cycle and is believed to act as a "molecular policeman," shutting down replication when the DNA is damaged or blocking the reproduction of DNA viruses (see Lane, 1992, *Nature* 358:15-16, incorporated herein by reference). The p53 transcription factor is part of a fundamental pathway which controls cell growth. Wild-type p53 can halt cell growth, or in some cases bring about programmed cell death (apoptosis). Such tumor-suppressive effects are absent in a variety of known p53 gene mutations. Moreover, p53 mutants not only deprive a cell of wild-type p53 tumor suppression, they also may spur abnormal cell growth.

In tumor cells, p53 is the most commonly mutated gene discovered to date (see Levine et al., 1991, *Nature* 351:453-456, and Hollstein et al., 1991, *Science* 253:49-53, each of which is incorporated herein by reference). Over half of the 6.5 million patients diagnosed with cancer annually possess p53 mutations in their tumor cells. Among common tumors, about 70% of colorectal cancers, 50% of lung cancers and 40% of breast cancers contain p53 mutations. In all, over 51 types of human tumors have been documented to possess p53 mutations, including bladder, brain, breast, cervix, colon, esophagus, larynx, liver, lung, ovary, pancreas, prostate, skin, stomach, and thyroid tumors (Culotta & Koshland, *Science* 262, 1958-1961 (1993); Rodrigues et al., 1990, *PNAS* 87:7555-7559, incorporated herein by reference). According to data presented by David Sidransky (1992 San Diego Conference), over 400 mutations in p53 are known. The presence of a p53 mutation in a tumor has also been correlated with a patient's prognosis. Patients who possess p53 mutations have a lower 5-year survival rate.

Proper diagnosis of the form of p53 in tumor cells is critical to clinicians to prescribe appropriate therapeutic regimens. For instance, patients with breast cancer who show no invasion of nearby lymph nodes generally do not relapse

after standard surgical treatment and chemotherapy. Of the 25% who do relapse after surgery and chemotherapy, additional chemotherapy is appropriate. At present, there is no clear way to determine which patients will benefit from such additional chemotherapy prior to relapse. However, correlating p53 mutations to tumorigenicity and metastasis provides clinicians with a means to determine whether such additional treatments are warranted.

In addition to facilitating conventional chemotherapy, appropriate diagnosis of p53 mutations provides clinicians with the ability to identify individuals who will benefit the most from gene therapy techniques, in which appropriately operative p53 copies are restored to a tumor site. Clinical p53 gene therapy trials are presently underway (Culotta & Koshland, *supra*).

The analysis of p53 mutations can also be used to identify which carcinogens lead to particular tumors (Harris, *Science* 262, 1980-1981 (1993)). For instance, dietary aflatoxin B₁ exposure is associated with G:C to T:A transversions at residue 249 of p53 in hepatocellular carcinomas (Hsu et al., *Nature* 350, 427 (1991); Bressac et al., *Nature* 350, 429 (1991); Harris, *supra*).

While most described p53 mutations are somatic in origin, some types of cancer are associated with germline p53 mutation. For instance, Li-Fraumeni syndrome is a hereditary condition in which individuals receive mutant p53 alleles, resulting in the early onset of various cancers (Harris, *supra*); Frebourg et al., *PNAS* 89, 6413-6417 (1992); Malkin et al., *Science* 250, 1233 (1990)). These mutations are associated with instability in the rest of the genome, creating multiple genetic alterations, and eventually leading to cancer.

hMLH1 and hMSH2 are mismatch repair genes which are causal agents in hereditary nonpolyposis colorectal cancer in individuals with mutant hMLH1 or hMSH2 alleles (Fishel et al., *supra*, and Papadopoulos et al., *supra*). Hereditary nonpolyposis colorectal cancer is a common genetic disorders, affecting about 1 in 200 individuals (Lynch et al.,

Gastroenterology 104, 1535 (1993)). Detection of hMLH1 and hMSH2 mutations in the population allows diagnosis of nonpolyposis colorectal cancer prone individuals prior to the manifestation of disease. This allows for the implementation of special screening programs for cancer-prone individuals to ensure early detection of cancer, thereby enhancing survival rates of afflicted individuals. In addition, genetic counselors may use the information derived from hMLH1 and hMSH2 chips to improve family planning as described for cystic fibrosis chips. The detection of mutations in hMLH1 and hMSH2 individually or in combination with p53 can also be used by clinicians to assess cancer prognosis and treatment modality. Finally, the information can be used to target appropriate individuals for gene therapy.

The entire hMLH1 gene is less than 85 kbp in length, comprising 2268 coding nucleotides (Papadopoulos et al., *supra*). Sequences from the gene have been deposited with GenBank (accession number U07418). Mutations associated with hereditary nonpolyposis colorectal cancer include the deletion of exon 5 (codons 578-632), a 4 base pair deletion of codons 727 and 728 resulting in a shift in the reading frame of the gene, a 4 base pair insertion at codons 755 and 756 resulting in an extension of the COOH terminus, a 371 base pair deletion and frameshift mutation at position 347, and a transversion causing an alteration of codon 252 resulting in the insertion of a stop codon (*id.*).

hMSH2 is a human homologue of the bacterial MutS and *S. cerevisiae* MSH mismatch-repair genes. MSH2, like hMLH1 is associated with hereditary nonpolyposis cancer. Although only a few MSH2 gene samples from tumor tissue have been characterized, at least some tumor samples show a T to C transition mutation at position 2020 of the cDNA sequence, resulting in the loss of an intron-exon splice acceptor site.

In view of the role of mutations in p53, MSH2 and/or hMLH1 in hereditary predisposition to cancer, to neoplastic transformation events leading to cancer and to cancer prognosis, it is important to screen individuals to determine whether they possess mutant alleles, and to identify precisely

which mutations the individuals possess. Because many mutations are point mutations, or extremely small insertions or deletions, which are generally undetectable by standard Southern analysis, accurate diagnosis requires a capacity to
5 examine a gene nucleotide-by-nucleotide.

Mutations in the hMSH2, hMLH1 or p53 genes, irrespective of whether previously characterized, can be detected by any of the tiling strategies noted above. Reference sequences of interest include full-length genomic and cDNA sequences of
10 each of these genes and subsequences thereof, such as exons and introns. For example, each nucleotide in the 20 kb p53 genomic sequence can be tiled using the basic strategy with an array of about 80,000 probes. As in the CFTR chip, some reference sequences are comparatively short sequences
15 including the site of a known mutation and a few flanking nucleotides. Some chips tile reference sequences that encompass mutational "hot spots." For instance, a variety of cellular and oncoviral proteins bind to specific regions of p53, including Mdm2, SV40 T antigen, E1b from adenovirus and
20 E6 from human papilloma virus. These binding sites correlate to some extent with observed high frequency somatic mutation regions of p53 found in tumor cells from cancer patients (see Harris et al., supra). Hot spots include exons 2, 3, 5, 6, 7 and 8 and the intronic regions between exons 2 and 3, 3 and 4
25 and 4 and 5. Fragments of the hMLH1 gene of particular interest include those encoding codons 578-632, 727, 728, 347, 252. Some chips are tiled to read mutations in each of the hMSH2, hMLH1 and p53 genes, both wildtype and mutant versions.

Standard or asymmetric PCR can be used to generate the
30 target DNA used in the tiling assays described above. In general, PCR is used to amplify hMSH2, hMLH1 or p53 sequences from a tissue of interest such as a tumor. Mixed PCR reactions can also be used to generate hMSH2, hMLH1 or p53
sequences simultaneously in a single reaction mixture. Any of
35 the coding or noncoding sequences from the genes may be amplified for use in the block tiling assays described above.

Table 8 below provides examples of primers which are useful in synthesizing specific regions of hMSH2, hMLH1 and

p53. Other primers can readily be devised from the known genomic and cDNA sequences of the genes. The primers described in Table 8 specific for p53 amplification have ends tailored to facilitate cloning into standard restriction enzyme cloning sites.

Table 8: Examples of PCR primers useful in amplifying regions of p53, hMHH1 and hMSH2.

Region Amplified	Primer Sequence	Description
Exon 5 (p53)	TAA TAC GAC TCA CTA TAG GGA GA CCC TGG GCA ACC AGC CCT GTC GT	Exon 5 T7 Primer (5' T7 to p53 3').
Exon 5 (p53)	ATG CAA TTA ACC CTC ACT AAA GGG AGA CAC TTG TGC CCT GAC TTT CAA C	Exon 5 T3 Primer (5' T3 to p53 3').
Exon 6 (p53)	TAA TAC GAC TCA CTA TAG GGA GCC TCC TCC CAG AGA CCC	Exon 6 T7 Primer (5'T7 to p53 3').
Exon 6 (p53)	ATG CAA TTA ACC CTC ACT AA GGG AGA TCC CCA GGC CTC TGA TTC CTC ACT G	Exon 6 T3 Primer (5'T3 to p53 3').
Exon 7 (p53)	TAA TAC GAC TCA CTA TAG GGA CTG GGG CAC AGC CAG GCC AGT GTG CA	Exon 7 T7 Primer (5' T7 to p53 3').
Exon 7 (p53)	ATG CAA TTA ACC CTC ACT AAA GGG AGA GTC TCC CCA AGG CGC ACT GGC CTC A	Exon 7 T3 Primer (5' T3 to p53 3').
Exon 8 (p53)	TAA TAC GAC TCA CTA TAG GGA GGG CAT AAC TGC ACC CTT GGT CTC CTC C	Exon 8 T7 Primer (5' T7 to p53 3').
Exon 8 (p53)	ATG CAA TTA ACC CTC ACT AAA GGG AGA GGA CCT GAT TTC CTT ACT GCC TCT TGC	Exon 8 T3 Primer (5' T3 to p53 3').
hMSH2	GAC ATG GCG GTG CAG CCG AAG GAG A	Primer for MSH2, 5' to 3'. If used with MSH2 primer below, a 3033 base pair amplicon will result
hMSH2	CTA TGT CAA TTG CAA ACA GTG CTC AGT TAC AG	Primer for hMSH2 5'to 3'.
hMLH1	CTT GGC TCT TCT GGC GCC AAA ATG TCG TTC	Primer for hMLH1, 5'to 3'. If used with hMLH1 primer below, a 2484 base pair amplicon will result.
hMLH1	TAT GTT AAG ACA CAT CTA TTT ATT TAT AAT CAA TCC	Primer for hMLH1 5' to 3'.

After PCR amplification of the target amplicon one strand of the amplicon can be isolated, i.e., using a biotinylated primer that allows capture of the undesired strand on streptavidin beads. Alternatively, asymmetric PCR can be used to generate a single-stranded target. Another approach involves the generation of single stranded RNA from the PCR product by incorporating a T7 or other RNA polymerase promoter in one of the primers. The single-stranded material can optionally be fragmented to generate smaller nucleic acids with less significant secondary structure than longer nucleic acids.

In one such method, fragmentation is combined with labeling. To illustrate, degenerate 8-mers or other degenerate short oligonucleotides are hybridized to the single-stranded target material. In the next step, a DNA polymerase is added with the four different dideoxynucleotides, each labeled with a different fluorophore. Fluorophore-labeled dideoxynucleotide are available from a variety of commercial suppliers. Hybridized 8-mers are extended by a labeled dideoxynucleotide. After an optional purification step, i.e., with a size exclusion column, the labeled 9-mers are hybridized to the chip. Other methods of target fragmentation can be employed. The single-stranded DNA can be fragmented by partial degradation with a DNase or partial depurination with acid. Labeling can be accomplished in a separate step, i.e., fluorophore-labeled nucleotides are incorporated before the fragmentation step or a DNA binding fluorophore, such as ethidium homodimer, is attached to the target after fragmentation.

30

Exemplary Chips

a. Exon VI Chip

To illustrate the value of the DNA chips of the present invention in such a method, a DNA chip was synthesized by the VLSIPS™ method to provide an array of overlapping probes which represent or tile across a 60 base region of exon 6 of the p53 gene. To demonstrate the ability to detect substitution mutations in the target, twelve different single substitution

mutations (wild type and three different substitutions at each of three positions) were represented on the chip along with the wild type. Each of these mutations was represented by a series of twelve 12-mer oligonucleotide probes, which were

5 complementary to the wild type target except at the one substituted base. Each of the twelve probes was complementary to a different region of the target and contained the mutated base at a different position, e.g., if the substitution was at base 32, the set of probes would be complementary--with the

10 exception of base 32--to regions of the target 21-32, 22-33, and 32-43). This enabled investigation of the effect of the substitution position within the probe. The alignment of some of the probes with a 12-mer model target nucleic acid is shown in Figure 27.

15 To demonstrate the effect of probe length, an additional series of ten 10-mer probes was included for each mutation (see Figure 28). In the vicinity of the substituted positions, the wild-type sequence was represented by every possible overlapping 12-mer and 10-mer probe. To simplify

20 comparisons, the probes corresponding to each varied position were arranged on the chip in the rectangular regions with the following structure: each row of cells represents one substitution, with the top row representing the wild type. Each column contains probes complementary to the same region

25 of the target, with probes complementary to the 3'-end of the target on the left and probes complementary to the 5'-end of the target on the right. The difference between two adjacent columns is a single base shift in the positioning of the probes. Whenever possible, the series of 10-mer probes were

30 placed in four rows immediately underneath and aligned with the 4 rows of 12-mer probes for the same mutation.

To provide model targets, 5' fluoresceinated 12-mers containing all possible substitutions in the first position of codon 192 were synthesized (see the starred position in the

35 target in Figure 27). Solutions containing 10 nM target DNA in 6X SSPE, 0.25% Triton X-100 were hybridized to the chip at room temperature for several hours. While target nucleic was hybridized to the chip, the fluorophores on the chip were

excited by light from an argon laser, and the chip was scanned with an autofocus confocal microscope. The emitted signals were processed by a PC to produce an image using image analysis software. By 1 to 3 hours, the signal had reached a plateau; to remove the hybridized target and allow hybridization to another target, the chip was stripped with 60% formamide, 2 X SSPE at 17 °C for 5 minutes. The washing buffer and temperature can vary, but the buffer typically contains 2-to-3X SSPE, 10-to-60% formamide (one can use multiple washes, increasing the formamide concentration by 10% each wash, and scanning between washes to determine when the wash is complete), and optionally a small percentage of Triton X-100, and the temperature is typically in the range of 15-to-18°C

Very distinct patterns were observed after hybridization with targets with 1 base substitutions and visualization with a confocal microscope and software analysis, as shown in Figure 29. In general, the probes which form perfect matches with the target retain the highest signal. For example, in the first image, the 12-mer probes that form perfect matches with the wild-type (WT) target are in the first row (top). The 12-mer probes with single base mismatches are located in the second, third, and fourth rows and have much lower signals. The data is also depicted graphically in Figure 30. On each graph, the X ordinate is the position of the probe in its row on the chip, and the Y ordinate is the signal at that probe site after hybridization. When a target with a different one base substitution is hybridized the complementary set of probes has the highest signal (see pictures 2, 3, and 4 in Figure 29 and graphs 2, 3, and 4 in Figure 30). In each case, the probe set with no mismatches with the target has the highest signals. Within a 12-mer probe set, the signal was highest at position 6 or 7. The graphs show that the signal difference between 12-mer probes at the same X ordinate tended to be greatest at positions 5 and 8 when the target and the complementary probes formed 10 base pairs and 11 base pairs, respectively. Because tumors often have both WT and mutant p53 genes, mixed target

populations were also hybridized to the chip, as shown in Figure 31. When the hybridization solution consisted of a 1:1 mixture of WT 12-mer and a 12-mer with a substitution in position 7 of the target, the sets of probes that were perfectly matched to both targets showed higher signals than the other probe sets.

The hybridization efficiency of a 10-mer probe array as compared to a 12-mer probe array was also compared. The 10-mer and 12-mer probe arrays gave comparable signals (see graphs 1-4 in Figure 30 and graphs 1-4 in Figure 32). However, the 10-mer probe sets, which are in rows 5-8 (see images in Figure 29), seemed to be better in this model system than the 12-mer probe sets at resolving one target from another, consistent with the expectation that one base mismatches are more destabilizing for 10-mers than 12-mers. Hybridization results within probe sets perfectly matched to target also followed the expectation that, the more matches the individual probe formed with the target, the higher the signal. However, duplexes with two 3' dangles (see Figure 30, position 6 in graphs 1-4) have about as much signal as the probes which are matched along their entire length (see Figure 30, position 7, in graphs 1-4).

This illustrative model system shows that 12-mer targets that differ by one base substitutions can be readily distinguished from one another by the novel probe array provided by the invention and that resolution of the different 12-mer targets was somewhat better with the 10-mer probe sets than with the 12-mer probe sets.

b. Exon V Chip

To analyze DNA from exon 5 of the p53 tumor suppressor gene, a set of overlapping 17-mer probes was synthesized on a chip. The probes for the WT allele were synthesized so as to tile across the entire exon with single base overlaps between probes. For each WT probe, a sets of 4 additional probes, one for each possible base substitution at position 7, were synthesized and placed in a column relative to the WT probe. Exon 5 DNA was amplified by PCR with primers flanking the exon. One of the primers was labeled with fluorescein; the

other primer was labeled with biotin. After amplification, the biotinylated strand was removed by binding to streptavidin beads. The fluoresceinated strand was used in hybridization.

5 About 1/3 of the amplified, single-stranded nucleic acid was hybridized overnight in 5 X SSPE at 60°C to the probe chip (under a cover slip). After washing with 6 X SSPE, the chip was scanned using confocal microscopy. Figure 33 shows an image of the p53 chip hybridized to the target DNA. Analysis
10 of the intensity data showed that 93.5% of the 184 bases of exon 5 were called in agreement with the WT sequence (see Buchman et al., 1988, Gene 70: 245-252, incorporated herein by reference). The miscalled bases were from positions where probe signal intensities were tied (1.6%) and where non-WT
15 probes had the highest signal intensity (4.9%). Figure 34 illustrates how the actual sequence was read. Gaps in the sequence of letters in the WT rows correspond to control probes or sites. Positions at which bases are miscalled are represented by letters in italic type in cells corresponding
20 to probes in which the WT bases have been substituted by other bases.

As the diagram indicates, the miscalled bases are from the low intensity areas of the image, which may be due to secondary structure in the target or probes preventing
25 intermolecular hybridization. To diminish the effects due to secondary structure, one can employ shorter targets (i.e., by target fragmentation) or use more stringent hybridization conditions. In addition, the use of a set of probes synthesized by tiling across the other strand of a duplex
30 target can also provide sequence information buried in secondary structure in the other strand. It should be appreciated, however, that the pattern of low intensity areas that forms as a result of secondary structure in the target itself provides a means to identify that a specific target
35 sequence is present in a sample. Other factors that may contribute to lower signal intensities include differences in probe densities and hybridization stabilities.

These results demonstrate the advantages provided by the DNA chips of the invention to genetic analysis. As another example, heterozygous mutations are currently sequenced by an arduous process involving cloning and repurification of DNA.

5 The cloning step is required, because the gel sequencing systems are poor at resolving even a 1:1 mixture of DNA. First, the target DNA is amplified by PCR with primers allowing easy ligation into a vector, which is taken up by transformation of E. coli, which in turn must be cultured, typically on plates overnight. After growth of the bacteria, DNA is purified in a procedure that typically takes about 2 hours; then, the sequencing reactions are performed, which takes at least another hour, and the samples are run on the gel for several hours, the duration depending on the length of the fragment to be sequenced. By contrast, the present invention provides direct analysis of the PCR amplified material after brief transcription and fragmentation steps, saving days of time and labor.

20 D. Mitochondrial Genome Chips

A human cell may have several hundred mitochondria, each with more than one copy of mtDNA. There is strand asymmetry in the base compositions, with one strand (Heavy) being relatively G rich, and the other strand (Light) being C rich. The L strand is 30.9% A, 31.2% C, 13.1% G, and 24.7% T. Human mtDNA is information-rich, encoding some 22 tRNAs, 12S and 16S rRNAs, and 13 polypeptides involved in oxidative phosphorylation. No introns have been detected. RNAs are processed by cleavage at tRNA sequences, and polyadenylated postranscriptionally. In some transcripts, polyadenylation also creates the stop codon, illustrating the parsimony of coding. In many individuals, mtDNA can be treated as haploid. However, some individuals are heteroplasmic (have more than one mtDNA sequence), and the degree of heteroplasmy can vary from tissue to tissue. Also, the rate of replication of mtDNAs can differ and together with random segregation during cell division, can lead to changes in heteroplasmy over time.

The human mitochondrial genome is 16,569 nucleotides

long. The sequence of the L-strand is numbered arbitrarily from the MboI-5/7 boundary in the D-loop region. The complete sequence of the human mitochondrial genome has been published. See Anderson et al., *Nature* 290, 457-465 (1981).

- 5 Mitochondrial DNA is maternally inherited, and has a mutation rate estimated to be tenfold higher than single copy nuclear DNA (Brown et al., *Proc. Natl. Acad. Sci. USA* 76, 1967-1971 (1979)). Human mtDNAs differ, on average, by about 70 base substitutions (Wallace, *Ann. Rev. Biochem.* 61, 1175-1212 (1992)). Over 80% of substitutions are transitions (i.e., pyrimidine-pyrimidine or purine-purine).

- Analysis of mitochondrial DNA serves several purposes. Detection of mutations in the mitochondrial genome allows diagnosis of a number of diseases. The mitochondrial genome has been identified as the locus of several mutations associated with human diseases. Some of the mutations result in stop codons in structural genes. Such mutations have been mapped and associated with diseases, such as Leber's hereditary optic neuropathy, neurogenic muscular weakness, ataxia and retinitis pigmentosa. Other mutations (nucleotide substitutions) occur in tRNA coding sequences, and presumably cause conformational defects in transcribed tRNA molecules. Such mutations have also been mapped and associated with diseases such as Myoclonic Epilepsy and Ragged Red Fiber Disease. Another type of mutation commonly found is deletions and/or insertions. Some deletions span segments of several kb. Again, such mutations have been mapped and associated with diseases, for example, ocular myopathy and Person Syndrome. See Wallace, *Ann. Rev. Biochem.* 61-1175-1212 (1992) (incorporated by reference in its entirety for all purposes). Early detection of such diseases allows metabolic or genetic therapy to be administered before irretrievable damage has occurred. *Id.* Analysis of mitochondrial DNA is also important for forensic screening. Because the mitochondrial genome is a locus of high variability between individuals, sequencing a substantial length of mitochondrial DNA provides a fingerprint that is highly specific to an individual.

Analysis of mitochondrial DNA is also important for evolutionary and epidemiological studies.

The reference sequence can be an entire mitochondrial genome or any fragment thereof. For forensic and epidemiological studies, the reference sequence is often all or part of the D-loop region in which variability between individuals is greatest (e.g., from 16024-16401 and 29-408). For detection of mutations, analysis of the entire genome is useful as a reference sequence, but shorter segments including the sites of known mutations, and about 1-20 flanking bases are also useful. Some chips have probes tiling paired reference sequences, representing wildtype and mutant versions of a sequence. Tiling a second reference sequence is particularly useful for detecting an insertion mutation occurring in 30-50% of ocular myopathy and Pearson syndrome patients, which consists of direct repeats of the sequence ACCTCCCTCACCA. Some chips include reference sequences from more than one mitochondrial genome.

Mitochondrial reference sequences can be tiled using any of the strategies noted above. The block tiling strategy is particularly useful for analyzing short reference sequences or known mutations. Either the block strategy or the basic strategy is suitable for analyzing long reference sequences. In many of the tiling strategies, it is possible to use fewer probes compared with the number used in other chips without significant loss of sequence information. As noted above, most point mutations in mitochondrial DNA are transitions, so for each wildtype nucleotide in a reference sequence, one of the three possible nucleotide substitutions is much more likely than the other two. Accordingly, in the basic tiling strategy, for example, a reference sequence can be tiled using only two probe sets. One probe set comprises a plurality of probes, each probe having a segment exactly complementary to the reference sequence. The second probe set comprises a corresponding probe for each probe in the first set. However, a probe from the second probe set differs from the corresponding probe from the first probe set in an interrogation position, in which the probe from the second

probe set includes the transition of the nucleotide present in that position in the probe from the first probe set.

Target mitochondrial DNA can be amplified, labelled and fragmented prior to hybridization using the same procedures as described for other chips. Use of at least two labelled nucleotides is desirable to achieve uniform labelling. Some exemplary primers are described below and other primers can be designed from the known sequence of mitochondrial DNA. Because mitochondrial DNA is present in multiple copies per cell, it can also be hybridized directly to a chip without prior amplification.

Exemplary Chips

The invention provides a DNA chip for analyzing sequences contained in a 1.3 kb fragment of human mitochondrial DNA from the "D-loop" region, the most polymorphic region of human mitochondrial DNA. One such chip comprises a set of 269 overlapping oligonucleotide probes of varying length in the range of 9-14 nucleotides with varying overlaps arranged in 600 x 600 micron features or synthesis sites in an array 1 cm x 1 cm in size. The probes on the chip are shown in columnar form below. An illustrative mitochondrial DNA chip of the invention comprises the following probes (X, Y coordinates are shown, followed by the sequence; "DL3" represents the 3'-end of the probe, which is covalently attached to the chip surface.)

0	0	DL3AGTGGGGTATTT	1	1	DL3GGTTGGTTTGGG
1	0	DL3GGGTATTAGTT	2	1	DL3TGGGGTTTCTAG
2	0	DL3TTAGTTTATCCAA	3	1	DL3GTTTCTAGTGGG
30	3	0 DL3ATCCAAACCAGG	4	1	DL3AGTGGGGGGTGT
4	0	DL3ACCAGGATCGGA	5	1	DL3GGGGTGTCAAAT
5	0	DL3CGTGTGTGTGTGG	6	1	DL3GTCAAATACATCG
6	0	DL3CGTGTGTGTGTGGC	7	1	DL3ACATCGAATGGAG
7	0	DL3TCGTGTGTGTGTGG	8	1	DL3CGAATGGAGGAG
35	8	0 DL3GTAGGATGGGTC	9	1	DL3GAGGAGTTTCGT
9	0	DL3AGGATGGGTCGT	10	1	DL3TTTTCGTTATGTGA
10	0	DL3GATGGGTCGTGT	11	1	DL3ATGTGACTTTTAC
11	0	DL3TGGCGACGATTG	12	1	DL3GACTTTTACAAAT
12	0	DL3GCGACGATTGGG	13	1	DL3AAATCTGCCCGA
40	13	0 DL3TGGGGGGGA	14	1	DL3AATCTGCCCGAG
14	0	DL3GAGGGGGCG	15	1	DL3CCCGAGTGTAGT
15	0	DL3GGAGGGGGCGA	16	1	DL3AGTGTAGTGGGG
16	0	DL3GAGGGGGCGA	0	2	DL3GGGAGGGTGAG
0	1	DL3GGCTTGTTGG	1	2	DL3GGTGAGGGTATG

2	2	DL3GGTATGATGATTAG	8	5	DL3ATTGTTAAACTTA
3	2	DL3GATTAGAGTAAGT	9	5	DL3AAACTTACAGACG
4	2	DL3TTAGAGTAAGTTA	10	5	DL3ACAGACGTGTCTG
5	2	DL3AAGTTATGTTGGG	11	5	DL3GTGTCTCGGTGAAA
5	6	DL3GTTGGGGGCG	12	5	DL3GTGAAAGGTGTGT
7	2	DL3GGGGCGGGTA	13	5	DL3GGTGTGTCTGTAG
8	2	DL3GCCGGTAGGAT	14	5	DL3TGTGTCTGTAGTA
9	2	DL3GGTAGGATGGGT	15	5	DL3GTAGTATTGTTTT
10	2	DL3GGATGGGTCGTG	16	5	DL3AGTATTGTTTTTT
10	11	DL3GGTCTGTGTGTGT	0	6	DL3CCTCGTGGGATA
12	2	DL3GTGTGTGTGGCG	1	6	DL3TGGGATACAGCG
13	2	DL3TGTGGCGACGAT	2	6	DL3GATACAGCGTCAT
14	2	DL3GACGATTGGGGT	3	6	DL3GCCGTCATAGACAG
15	2	DL3ATTGGGGTATGG	4	6	DL3AGACAGAACTAA
15	16	DL3GTATGGGGCTTG	5	6	DL3CAGAACTAAGGA
0	3	DL3GGATTGTGGTCG	6	6	DL3TAAGGACGGAGT
1	3	DL3TGGTCTGGATTGG	7	6	DL3GACGGAGTAGGA
2	3	DL3GGATTGGTCTAAA	8	6	DL3GTAGGATAATAAA
3	3	DL3TCTAAAGTTTAAA	9	6	DL3TAATAAATAGCG
20	4	DL3GTTTTAAAATAGAA	10	6	DL3ATAGCGTAGGAT
5	3	DL3ATAGAAAAACCG	11	6	DL3TAGCGTAGGATG
6	3	DL3AGAAAAACCGC	12	6	DL3AGGATGCAAGTT
7	3	DL3AACC GCCATAC	13	6	DL3ATGCAAGTTATAA
8	3	DL3CCATACGTGAAAA	14	6	DL3GTTATAATGTCCG
25	9	DL3ACGTGAAAAATTGT	15	6	DL3ATGTCCGCTTGT
10	3	DL3AATTGTCA GTGGG	16	6	DL3TCCGCTTGTATG
11	3	DL3TGTCA GTGGGGG	0	7	DL3GTGAGTGCCCTC
12	3	DL3TGGGGGGTTGA	1	7	DL3TGCCCTCGAGAG
13	3	DL3GGGTTGATTGTGT	2	7	DL3CCTCGAGAGGTA
30	14	DL3TTGTGTAATAAAA	3	7	DL3AGAGGTACGTAA
15	3	DL3AATAAAAAGGGGA	4	7	DL3ACGTAAACCATA
16	3	DL3TAAAAGGGGAGG	5	7	DL3ACCATAAAAAGCAG
0	4	DL3GTTTTTTTAAAGG	6	7	DL3AAAGCAGACCC
1	4	DL3TTTTTAAAGGTGG	7	7	DL3AGACCCCCCAT
35	2	DL3AGGTGGTTTGG	8	7	DL3CCCCCATACGT
3	4	DL3TTGGGGGGGAG	9	7	DL3CATACGTGCGCT
4	4	DL3GGAGGGGGGCG	10	7	DL3GTGCGCTATCAG
5	4	DL3GGGGCGAAGAC	11	7	DL3GCGCTATCAGTA
6	4	DL3GAAGACCGGATG	12	7	DL3TCAGTAACGCTC
40	7	DL3CCGGATGTCGTG	13	7	DL3GTAACGCTCTGC
8	4	DL3GTCTGTGAATTTGT	14	7	DL3CTCTGCGACCTC
9	4	DL3CGTGAATTTGTGT	15	7	DL3GACCTCGGCCT
10	4	DL3TTGTGTAGAGACG	16	7	DL3TCGGCCTCGTG
11	4	DL3TAGAGACGGTTT	0	8	DL3GATGAAGTCCCAG
45	12	DL3ACGGTTTGGGG	1	8	DL3AGTCCCAGTATTT
13	4	DL3TGGGGTTTTTGT	2	8	DL3GTATTTTCGGATTT
14	4	DL3GGGTTTTTGT	3	8	DL3TCGGATTTATCG
15	4	DL3TTGTTTCTTGGG	4	8	DL3GATTTATCGGGT
16	4	DL3TCTTGGGATTGTG	5	8	DL3ATCGGGTGTGCA
50	0	DL3TGTATGAATGATTT	6	8	DL3TGTGCAAGGGGA
1	5	DL3TGATTTACACAA	7	8	DL3CAAGGGGAATTT
2	5	DL3ACACAATTAATTAA	8	8	DL3GAATTTATTCTGTA
3	5	DL3AATTAATTACGAA	9	8	DL3TCTGTAGTGCTAC
4	5	DL3TACGAACATCCTG	10	8	DL3GTAGTGCTACCT
55	5	DL3ACGAACATCCTGT	11	8	DL3GCTACCTAGTAG
6	5	DL3TCCTGTATTATTA	12	8	DL3CTAGTAGTCCAGA
7	5	DL3GTATTATTATTGTT	13	8	DL3TCCAGATAGTGGG

14	8	DL3AGATAGTGGGATA	8	12	DL3TGTTTCGTTTCATGT
15	8	DL3GGGATAAATTGGT	9	12	DL3CGTTCATGTCGTT
16	8	DL3TAATTGGTGAGTG	10	12	DL3GTCGTTAGTTGG
0	9	DL3TATAGGGCGTGT	11	12	DL3TAGTTGGGAGTT
5	1	DL3GGCGTGTTCTCA	12	12	DL3GGAGTTGATAGTG
2	9	DL3GTGTTCTCACGAT	13	12	DL3ATAGTGTGTAGTT
3	9	DL3TCACGATGAGAGG	14	12	DL3GTGTAGTTGACGT
4	9	DL3ATGAGAGGAGCG	15	12	DL3TGACGTTGAGGT
5	9	DL3AGGAGCGAGGC	16	12	DL3CGTTGAGGTTTA
10	6	DL3CGAGGCCCGG	5	13	DL3TATAACATGCCAT
7	9	DL3GCCCGGGTATT	6	13	DL3AACATGCCATGGT
8	9	DL3CGGGTATTGTGA	7	13	DL3CCATGGTATTTAT
9	9	DL3GTGAACCCCAT	8	13	DL3ATTTATGAACTGG
10	9	DL3CCCCATCGATTT	9	13	DL3AACTGGTGGACAT
15	11	DL3ATCGATTTCACTT	10	13	DL3TGGACATCATGTA
12	9	DL3TTTCACTTGACAT	11	13	DL3CATGTATTTTTGG
13	9	DL3TTGACATAGAGCT	12	13	DL3TTTTGGGTTAGG
14	9	DL3TAGAGCTGTAGAC	13	13	DL3GGGTTAGGATGT
15	9	DL3GTAGACCAAGGA	14	13	DL3GGATGTAGTTTTG
20	16	DL3ACCAAGGATGAAG	15	13	DL3TGAGTTTTTGGG
0	10	DL3CGTGTAAATGTCAG	16	13	DL3TTTGGGGGAGG
1	10	DL3TGTCAGTTTAGGG	5	14	DL3GGGTTTCATAACTG
2	10	DL3TCAGTTTAGGGA	6	14	DL3ATAACTGAGTGGG
3	10	DL3TAGGGAAGAGCA	7	14	DL3AACTGAGTGGGT
25	4	DL3AAGAGCAGGGGT	8	14	DL3GTGGGTAGTTGT
5	10	DL3CAGGGGTACCTA	9	14	DL3GTAGTTGTTGGC
6	10	DL3GGTACCTACTGG	10	14	DL3GTTGGCGATACA
7	10	DL3TACTGGGGGGA	11	14	DL3CGATACATAAAAAG
8	10	DL3GGGGGAGTCTAT	12	14	DL3TAAAAGCATGTAA
30	9	DL3AGTCTATCCCCA	13	14	DL3GCATGTAATGACG
10	10	DL3ATCCCCAGGGA	14	14	DL3ATGACGGTTCGGT
11	10	DL3CAGGGAACTGGT	15	14	DL3GTCGGTGGTACT
12	10	DL3ACTGGTGGTAGG	16	14	DL3GGTACTTATAACA
13	10	DL3CTGGTGGTAGGA	5	15	DL3TCGATTCTAAGAT
35	14	DL3GTAGGAGGCACA	6	15	DL3TAAGATTAAATTT
15	10	DL3GGCACATTTAGT	7	15	DL3AAATTTGAATAAG
16	10	DL3TTTAGTTATAGGG	8	15	DL3AATAAGAGACAAG
0	11	DL3AGGTTTACGGTG	9	15	DL3AAGAGACAAGAAA
1	11	DL3TACGGTGGGGA	10	15	DL3AAGAAAGTACCC
40	2	DL3GTGGGGAGTGG	11	15	DL3AAAGTACCCCTT
3	11	DL3GGGAGTGGGTGA	12	15	DL3CCCCTTCGTCTA
4	11	DL3GGGTGATCCTATG	13	15	DL3CTTCGTCTAAAC
5	11	DL3CCTATGGTTGTTT	14	15	DL3CTAAACCCATGG
6	11	DL3GGTTGTTTGGATG	15	15	DL3AACCCATGGTGG
45	7	DL3GTTTGGATGGGT	16	15	DL3TGGTGGGTTTCAT
8	11	DL3ATGGGTGGGAAT	5	16	DL3TTGGAAAAAGGT
9	11	DL3GGGAATTGTCATG	6	16	DL3AAAAGGTTCTCTG
10	11	DL3GTCATGTATCATGT	7	16	DL3GGTTCCTGTTTA
11	11	DL3TCATGTATTTCCG	8	16	DL3CCTGTTTAGTCTC
50	12	DL3TATTTCCGTAAA	9	16	DL3TTAGTCTCTTTTT
13	11	DL3TTCCGTAAATGG	10	16	DL3CTTTTTTCAGAAAT
14	11	DL3GTAAATGGCATGT	11	16	DL3AGAAATTGAGGTG
15	11	DL3GCATGTAATCGTG	12	16	DL3AAATTGAGGTGGT
16	11	DL3GTAATCGTGTAAT	13	16	DL3GGTGGAATCGT
55	5	DL3GGGAGGGGTAC	14	16	DL3TAATCGTGGGTT
6	12	DL3GGGTACGAATGT	15	16	DL3GTGGGTTTCGAT
7	12	DL3ACGAATGTTTCGTT	16	16	DL3GGTTTCGATTCT

No probes were present in positions X, Y = 0, 12 to X, Y = 4, 12; X, Y = 0, 13 to X, Y = 4, 13; X, Y = 0, 14 to X, Y = 4, 14; X, Y = 0, 15 to X, Y = 4, 15; X, Y = 0, 16 to X, Y = 4, 16;

The length of each of the probes on the chip was variable to minimize differences in melting temperature and potential for cross-hybridization. Each position in the sequence was represented by at least one probe and most positions were represented by 2 or more probes. As noted above, the amount of overlap between the oligonucleotides varied from probe to probe. Figure 35 shows the human mitochondrial genome; "O_H" is the H strand origin of replication, and arrows indicate the cloned unshaded sequence.

DNA was prepared from hair roots of six human donors (mt1 to mt6) and then amplified by PCR and cloned into M13; the resulting clones were sequenced using chain terminators to verify that the desired specific sequences were present. DNA from the sequenced M13 clones was amplified by PCR, transcribed *in vitro*, and labeled with fluorescein-UTP using T3 RNA polymerase. The 1.3 kb RNA transcripts were fragmented and hybridized to the chip. The results showed that each different individual had DNA that produced a unique hybridization fingerprint on the chip and that the differences in the observed patterns could be correlated with differences in the cloned genomic DNA sequence. The results also demonstrated that very long sequences of a target nucleic acid can be represented comprehensively as a specific set of overlapping oligonucleotides and that arrays of such probe sets can be usefully applied to genetic analysis.

The sample nucleic acid was hybridized to the chip in a solution composed of 6 X SSPE, 0.1% Triton-X 100 for 60 minutes at 15°C. The chip was then scanned by confocal scanning fluorescence microscopy. The individual features on the chip were 588 x 588 microns, but the lower left 5 x 5 square features in the array did not contain probes. To quantitate the data, pixel counts were measured within each synthesis site. Pixels represent 50 x 50 microns. The fluorescence intensity for each feature was scaled to a mean

determined from 27 bright features. After scanning, the chip was stripped and rehybridized; all six samples were hybridized to the same chip. Figure 36 shows the image observed from the mt4 sample on the DNA chip. Figure 37 shows the image

5 observed from the mt5 sample on the DNA chip. Figure 38 shows the predicted difference image between the mt4 and mt5 samples on the DNA chip based on mismatches between the two samples and the reference sequence (see Anderson et al., *supra*). Figure 39 shows the actual difference image observed.

10 The results show that, in almost all cases, mismatched probe/target hybrids resulted in lower fluorescence intensity than perfectly matched hybrids. Nonetheless, some probes detected mutations (or specific sequences) better than others, and in several cases, the differences were within noise
15 levels. Improvements can be realized by increasing the amount of overlap between probes and hence overall probe density and, for duplex DNA targets, using a second set of probes, either on the same or a separate chip, corresponding to the second strand of the target. Figure 40, in sheets 1 and 2, shows a
20 plot of normalized intensities across rows 10 and 11 of the array and a tabulation of the mutations detected.

Figure 41 shows the discrimination between wild-type and mutant hybrids obtained with this chip. The median of the six normalized hybridization scores for each probe was taken. The
25 graph plots the ratio of the median score to the normalized hybridization score versus mean counts. On this graph, a ratio of 1.6 and mean counts above 50 yield no false positives, and while it is clear that detection of some mutants can be improved, excellent discrimination is achieved,
30 considering the small size of the array. Figure 42 illustrates how the identity of the base mismatch may influence the ability to discriminate mutant and wild-type sequences more than the position of the mismatch within an oligonucleotide probe. The mismatch position is expressed as
35 % of probe length from the 3'-end. The base change is indicated on the graph. These results show that the DNA chip increases the capacity of the standard reverse dot blot format by orders of magnitude, extending the power of that approach

many fold and that the methods of the invention are more efficient and easier to automate than gel-based methods of nucleic acid sequence and mutation analysis.

To illustrate further these advantages, a second chip was prepared for analyzing a longer segment from human mitochondrial DNA (mtDNA). The chip "tiles" through 648 nucleotides of a reference sequence comprising human H strand mtDNA from positions 16280 to 356, and allows analysis of each nucleotide in the reference sequence. The probes in the array are 15 nucleotides in length, and each position in the target sequence is represented by a set of 4 probes (A, C, G, T substitutions), which differed from one another at position 7 from the 3'-end. The array consists of 13 blocks of 4 x 50 probes: each block scans through 50 nucleotides of contiguous mtDNA sequence. The blocks are separated by blank rows. The 4 corner columns contain control probes; there are a total of 2600 probes in a 1.28 cm x 1.28 cm square area (feature), and each area is 256 x 197 microns.

Target RNA was prepared as above. The RNA was fragmented and hybridized to the oligonucleotide array in a solution composed of 6X SSPE, 0.1% Triton X-100 for 60 minutes at 18°C. Unhybridized material was washed away with buffer, and the chip was scanned at 25 micron pixel resolution.

Figure 43 provides a 5' to 3' sequence listing of one target corresponding to the probes on the chip. X is a control probe. Positions that differ in the target (i.e., are mismatched with the probe at the designated site) are in bold. Figure 44 shows the fluorescence image produced by scanning the chip when hybridized to this sample. About 95% of the sequence could be read correctly from only one strand of the original duplex target nucleic acid. Although some probes did not provide excellent discrimination and some probes did not appear to hybridize to the target efficiently, excellent results were achieved. The target sequence differed from the probe set at six positions: 4 transitions and 2 insertions. All 4 transitions were detected, and specific probes could readily be incorporated into the array to detect insertions or deletions. Figure 45 illustrates the detection of 4

transitions in the target sequence relative to the wild-type probes on the chip.

A further chip was constructed comprising probes tiling across the entire D-loop region (1.3 kb) of mt DNA sequences from two humans. The probes were tiled in rows of four using the basic tiling strategy. The probes were overlapping 15 mers having an interrogation position 7 nucleotides from the 3' end. The complete group of probes tiled on the reference sequence from the first individual, designated mt1, occupied the upper half of the chip. The lower half of the chip contained a similar arrangement based on a second clone, mt2. The probes were synthesized in a 1.28 x 1.28 cm area, which contained a matrix of 115 x 120 cells. The chip contained a total of 10,488 mtDNA probes.

Six samples of target DNA was extracted from hair roots from six individuals. The 1.3 kb region spanning positions 15935 to 667 of human mtDNA was PCR amplified, cloned in bacteriophage M13 and sequenced by conventional methods. The 1.3 kb region was reamplified from the phage clone using primers, L15935-T3, 5'CTCGGAATTAACCCTCACTAAAGGAAACCTTTTCCAAGGA and H667-T7, 5'TAATACGACTCACTATAGGGAGAGGCTAGGACCAAACCTATT tagged with T3 and T7 RNA polymerase promoter sequences. Labelled RNA was generated by *in vitro* transcription using T3 RNA polymerase and fluoresceinated nucleotides, fragmented, and hybridized to the mtDNA control region resequencing chip at room temperature for 60 min, in 6xSSPE + 0.05% triton X-100. Six washes were carried out at room temperature, using 6xSSPE + 0.005% triton X-100, and the chip was read. Signal intensities varied considerably over the chip, but the large dynamic range of the detection system allowed accurate quantitation of intensities over several orders of magnitude. Even relatively low signal intensities yielded accurate results.

Five different clones (mt1-5) were hybridized, each to a separate chip. The reference sequence was also hybridized for comparative purposes. Mean counts per probe cell were determined, and used by automated basecalling software to read the sequence. The accuracy of sequence read from the chip is

summarized as follows. Combining the data from the five targets analyzed, the chip read a total of 6310 nucleotides. Of these nucleotides in the target sequences, 55 were different from the reference sequence (as judged by conventional sequencing) (41 of these 55 nucleotides were both detected and read correctly from the chip). 6 of 55 nucleotides were detected as being ambiguous but their identity could not be read. 2 of 55 nucleotides were detected as mutations, but their identity was miscalled. 6 of 55 nucleotides were incorrectly called as wildtype. Of the 6255 nucleotides in the target sequence that were identical to the reference sequence, only 36 (0.57%) were miscalled or scored as ambiguous.

A further chip was constructed comprising probes tiling across a reference sequence comprising an entire mitochondrial genome. In this chip, a block tiling strategy was used. Each block was designed to analyze seven nucleotides from a target sequence. Each block consisted of four probe sets, the probe sets each having seven probes. A block was laid down on the chip in seven columns of four probes. The upper probe was the same in each column, this being a probe exactly complementary to a subsequence of the reference sequence. The three other probes in each column were identical to the upper probe except in an interrogation position, which was occupied by a different base in each of the four probes in the column. The interrogation position shifted by one position between successive columns. Thus, except for the seven interrogation positions, one in each of the columns of probes, all probes occupying a block were identical. The array comprised many such blocks, each tiled to successive subsequences of the mitochondrial DNA reference sequence. In all, the chip tiled 15,569 nucleotides of reference sequence with double tiling at 42 positions. 66,276 probes occupied an array of 304 x 315 cells, each cell having an area of 42 x 41 microns.

The chip was hybridized to the same target sequences as described for the D-loop region, except that hybridization was at 15°C for 2 hr. The chip was scanned at 5 micron resolution to give an image with approximately 64 pixels per cell. For

blocks of probes tiling across the D-loop region, a sequence-specific hybridization pattern was obtained. For other blocks, only background hybridization was observed.

These results illustrate that longer sequences can be read using the DNA chips and methods of the invention, as compared to conventional sequencing methods, where reading length is limited by the resolution of gel electrophoresis. Hybridization and signal detection require less than an hour and can be readily shortened by appropriate choice of buffers, temperatures, probes, and reagents.

III. MODES OF PRACTICING THE INVENTION

A. VLSIPS™ Technology

As noted above, the VLSIPS™ technology is described in a number of patent publications and is preferred for making the oligonucleotide arrays of the invention. A brief description of how this technology can be used to make and screen DNA chips is provided in this Example and the accompanying Figures. In the VLSIPS™ method, light is shone through a mask to activate functional (for oligonucleotides, typically an -OH) groups protected with a photoremovable protecting group on a surface of a solid support. After light activation, a nucleoside building block, itself protected with a photoremovable protecting group (at the 5'-OH), is coupled to the activated areas of the support. The process can be repeated, using different masks or mask orientations and building blocks, to prepare very dense arrays of many different oligonucleotide probes. The process is illustrated in Figure 46; Figure 47 illustrates how the process can be used to prepare "nucleoside combinatorials" or oligonucleotides synthesized by coupling all four nucleosides to form dimers, trimers and so forth.

New methods for the combinatorial chemical synthesis of peptide, polycarbamate, and oligonucleotide arrays have recently been reported (see Fodor et al., 1991, *Science* 251: 767-773; Cho et al., 1993, *Science* 261: 1303-1305; and Southern et al., 1992, *Genomics* 13: 1008-10017, each of which is incorporated herein by reference). These arrays, or

biological chips (see Fodor et al., 1993, *Nature* 364: 555-556, incorporated herein by reference), harbor specific chemical compounds at precise locations in a high-density, information rich format, and are a powerful tool for the study of biological recognition processes. A particularly exciting application of the array technology is in the field of DNA sequence analysis. The hybridization pattern of a DNA target to an array of shorter oligonucleotide probes is used to gain primary structure information of the DNA target. This format has important applications in sequencing by hybridization, DNA diagnostics and in elucidating the thermodynamic parameters affecting nucleic acid recognition.

Conventional DNA sequencing technology is a laborious procedure requiring electrophoretic size separation of labeled DNA fragments. An alternative approach, termed Sequencing By Hybridization (SBH), has been proposed (Lysov et al., 1988, *Dokl. Akad. Nauk SSSR* 303:1508-1511; Bains et al., 1988, *J. Theor. Biol.* 135:303-307; and Drmanac et al., 1989, *Genomics* 4:114-128, incorporated herein by reference). This method uses a set of short oligonucleotide probes of defined sequence to search for complementary sequences on a longer target strand of DNA. The hybridization pattern is used to reconstruct the target DNA sequence. It is envisioned that hybridization analysis of large numbers of probes can be used to sequence long stretches of DNA. In immediate applications of this hybridization methodology, a small number of probes can be used to interrogate local DNA sequence.

The strategy of SBH can be illustrated by the following example. A 12-mer target DNA sequence, AGCCTAGCTGAA, is mixed with a complete set of octanucleotide probes. If only perfect complementarity is considered, five of the 65,536 octamer probes -TCGGATCG, CGGATCGA, GGATCGAC, GATCGACT, and ATCGACTT will hybridize to the target. Alignment of the overlapping sequences from the hybridizing probes reconstructs the complement of the original 12-mer target:

```
TCGGATCG
CGGATCGA
GGATCGAC
GATCGACT
```

ATCGACTT
TCGGATCGACTT

Hybridization methodology can be carried out by attaching
5 target DNA to a surface. The target is interrogated with a
set of oligonucleotide probes, one at a time (see Strezoska et
al., 1991, *Proc. Natl. Acad. Sci. USA* 88:10089-10093, and
Drmanac et al., 1993, *Science* 260:1649-1652, each of which is
incorporated herein by reference). This approach can be
10 implemented with well established methods of immobilization
and hybridization detection, but involves a large number of
manipulations. For example, to probe a sequence utilizing a
full set of octanucleotides, tens of thousands of
hybridization reactions must be performed. Alternatively, SBH
15 can be carried out by attaching probes to a surface in an
array format where the identity of the probes at each site is
known. The target DNA is then added to the array of probes.
The hybridization pattern determined in a single experiment
directly reveals the identity of all complementary probes.

20 As noted above, a preferred method of oligonucleotide
probe array synthesis involves the use of light to direct the
synthesis of oligonucleotide probes in high-density,
miniaturized arrays. Photolabile 5'-protected
N-acyl-deoxynucleoside phosphoramidites, surface linker
25 chemistry, and versatile combinatorial synthesis strategies
have been developed for this technology. Matrices of
spatially-defined oligonucleotide probes have been generated,
and the ability to use these arrays to identify complementary
sequences has been demonstrated by hybridizing fluorescent
30 labeled oligonucleotides to the DNA chips produced by the
methods. The hybridization pattern demonstrates a high degree
of base specificity and reveals the sequence of
oligonucleotide targets.

The basic strategy for light-directed oligonucleotide
35 synthesis (1) is outlined in Fig. 46. The surface of a solid
support modified with photolabile protecting groups (X) is
illuminated through a photolithographic mask, yielding
reactive hydroxyl groups in the illuminated regions. A
3'-O-phosphoramidite activated deoxynucleoside (protected at

the 5'-hydroxyl with a photolabile group) is then presented to the surface and coupling occurs at sites that were exposed to light. Following capping, and oxidation, the substrate is rinsed and the surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of products is obtained.

10 Light directed chemical synthesis lends itself to highly efficient synthesis strategies which will generate a maximum number of compounds in a minimum number of chemical steps. For example, the complete set of 4^n polynucleotides (length n), or any subset of this set can be produced in only $4 \times n$ chemical steps. See Fig. 47. The patterns of illumination and the order of chemical reactants ultimately define the products and their locations. Because photolithography is used, the process can be miniaturized to generate high-density arrays of oligonucleotide probes. For an example of the

20 nomenclature useful for describing such arrays, an array containing all possible octanucleotides of dA and dT is written as $(A+T)^8$. Expansion of this polynomial reveals the identity of all 256 octanucleotide probes from AAAAAAAAAA to TTTTTTTT. A DNA array composed of complete sets of

25 dinucleotides is referred to as having a complexity of 2. The array given by $(A+T+C+G)^8$ is the full 65,536 octanucleotide array of complexity four. Computer-aided methods of laying down predesigned arrays of probes using VLSIPS™ technology are described in commonly-assigned co-pending application USSN

30 08/249,188, filed May 24, 1994 (incorporated by reference in its entirety for all purposes).

To carry out hybridization of DNA targets to the probe arrays, the arrays are mounted in a thermostatically controlled hybridization chamber. Fluorescein labeled DNA

35 targets are injected into the chamber and hybridization is allowed to proceed for 5 min to 24 hr. The surface of the matrix is scanned in an epifluorescence microscope (Zeiss Axioscop 20) equipped with photon counting electronics using

50 - 100 μ W of 488 nm excitation from an Argon ion laser (Spectra Physics Model 2020). Measurements may be made with the target solution in contact with the probe matrix or after washing. Photon counts are stored and image files are presented after conversion to an eight bit image format. See Fig. 51.

When hybridizing a DNA target to an oligonucleotide array, $N = L_t - (L_p - 1)$ complementary hybrids are expected, where N is the number of hybrids, L_t is the length of the DNA target, and L_p is the length of the oligonucleotide probes on the array. For example, for an 11-mer target hybridized to an octanucleotide array, $N = 4$. Hybridizations with mismatches at positions that are 2 to 3 residues from either end of the probes will generate detectable signals. Modifying the above expression for N , one arrives at a relationship estimating the number of detectable hybridizations (N_d) for a DNA target of length L_t and an array of complexity C . Assuming an average of 5 positions giving signals above background:

$$N_d = (1 + 5(C-1))[L_t - (L_p - 1)].$$

Arrays of oligonucleotides can be efficiently generated by light-directed synthesis and can be used to determine the identity of DNA target sequences. Because combinatorial strategies are used, the number of compounds increases exponentially while the number of chemical coupling cycles increases only linearly. For example, synthesizing the complete set of 4^8 (65,536) octanucleotides will add only four hours to the synthesis for the 16 additional cycles. Furthermore, combinatorial synthesis strategies can be implemented to generate arrays of any desired composition. For example, because the entire set of dodecamers (4^{12}) can be produced in 48 photolysis and coupling cycles (b^n compounds requires $b \times n$ cycles), any subset of the dodecamers (including any subset of shorter oligonucleotides) can be constructed with the correct lithographic mask design in 48 or fewer chemical coupling steps. In addition, the number of compounds in an array is limited only by the density of synthesis sites and the overall array size. Recent experiments have demonstrated hybridization to probes

synthesized in 25 μm sites. At this resolution, the entire set of 65,536 octanucleotides can be placed in an array measuring 0.64 cm square, and the set of 1,048,576 dodecanucleotides requires only a 2.56 cm array.

5 Genome sequencing projects will ultimately be limited by DNA sequencing technologies. Current sequencing methodologies are highly reliant on complex procedures and require substantial manual effort. Sequencing by hybridization has the potential for transforming many of the manual efforts into
10 more efficient and automated formats. Light-directed synthesis is an efficient means for large scale production of miniaturized arrays for SBH. The oligonucleotide arrays are not limited to primary sequencing applications. Because single base changes cause multiple changes in the
15 hybridization pattern, the oligonucleotide arrays provide a powerful means to check the accuracy of previously elucidated DNA sequence, or to scan for changes within a sequence. In the case of octanucleotides, a single base change in the target DNA results in the loss of eight complements, and
20 generates eight new complements. Matching of hybridization patterns may be useful in resolving sequencing ambiguities from standard gel techniques, or for rapidly detecting DNA mutational events. The potentially very high information content of light-directed oligonucleotide arrays will change
25 genetic diagnostic testing. Sequence comparisons of hundreds to thousands of different genes will be assayed simultaneously instead of the current one, or few at a time format. Custom arrays can also be constructed to contain genetic markers for the rapid identification of a wide variety of pathogenic
30 organisms.

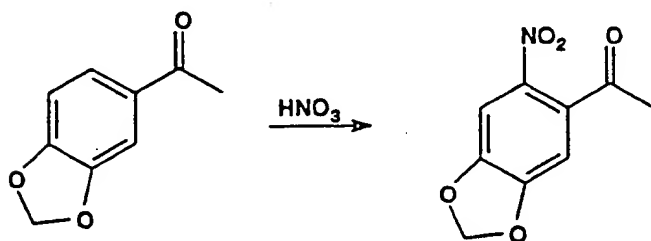
Oligonucleotide arrays can also be applied to study the sequence specificity of RNA or protein-DNA interactions.

Experiments can be designed to elucidate specificity rules of non Watson-Crick oligonucleotide structures or to investigate
35 the use of novel synthetic nucleoside analogs for antisense or triple helix applications. Suitably protected RNA monomers may be employed for RNA synthesis. The oligonucleotide arrays should find broad application deducing the thermodynamic and

kinetic rules governing formation and stability of oligonucleotide complexes.

Other than the use of photoremovable protecting groups, the nucleoside coupling chemistry is very similar to that used routinely today for oligonucleotide synthesis. Fig. 48 shows the deprotection, coupling, and oxidation steps of a solid phase DNA synthesis method. Fig. 49 shows an illustrative synthesis route for the nucleoside building blocks used in the method. Fig. 50 shows a preferred photoremovable protecting group, MeNPOC, and how to prepare the group in active form. The procedures described below show how to prepare these reagents. The nucleoside building blocks are 5'-MeNPOC-THYMIDINE-3'-OCEP; 5'-MeNPOC-N⁴-t-BUTYL PHENOXYACETYL-DEOXYCYTIDINE-3'-OCEP; 5'-MeNPOC-N⁴-t-BUTYL PHENOXYACETYL-DEOXYGUANOSINE-3'-OCEP; and 5'-MeNPOC-N⁴-t-BUTYL PHENOXYACETYL-DEOXYADENOSINE-3'-OCEP..

1. Preparation of 4,5-methylenedioxy-2-nitroacetophenone



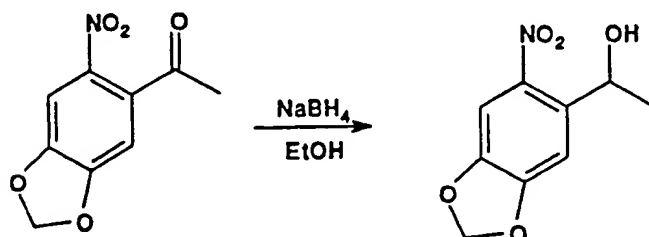
20

A solution of 50 g (0.305 mole) 3,4-methylenedioxyacetophenone (Aldrich) in 200 mL glacial acetic acid was added dropwise over 30 minutes to 700 mL of cold (2-4°C) 70% HNO₃ with stirring (NOTE: the reaction will overheat without external cooling from an ice bath, which can be dangerous and lead to side products). At temperatures below 0°C, however, the reaction can be sluggish. A temperature of 3-5°C seems to be optimal). The mixture was left stirring for another 60 minutes at 3-5°C, and then allowed to approach ambient temperature. Analysis by TLC (25% EtOAc in hexane) indicated complete conversion of the starting material within 1-2 hr. When the reaction was complete, the mixture was poured into 3 liters of crushed ice, and the resulting yellow solid was

filtered off, washed with water and then suction-dried. Yield ~53 g (84%), used without further purification.

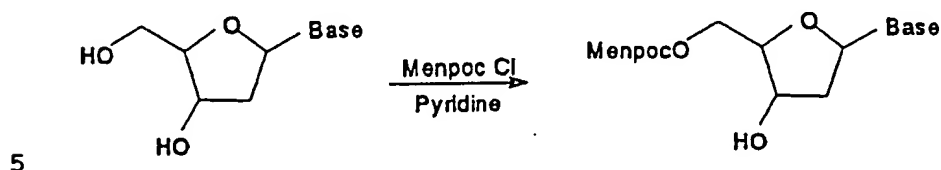
2. Preparation of 1-(4,5-Methylenedioxy-2-nitrophenyl)

5 ethanol



Sodium borohydride (10g; 0.27 mol) was added slowly to a cold, stirring suspension of 53g (0.25 mol) of 4,5-methylenedioxy-2-nitroacetophenone in 400 mL methanol. The temperature was kept below 10°C by slow addition of the NaBH_4 and external cooling with an ice bath. Stirring was continued at ambient temperature for another two hours, at which time TLC (CH_2Cl_2) indicated complete conversion of the ketone. The mixture was poured into one liter of ice-water and the resulting suspension was neutralized with ammonium chloride and then extracted three times with 400 mL CH_2Cl_2 or EtOAc (the product can be collected by filtration and washed at this point, but it is somewhat soluble in water and this results in a yield of only ~60%). The combined organic extracts were washed with brine, then dried with MgSO_4 and evaporated. The crude product was purified from the main byproduct by dissolving it in a minimum volume of CH_2Cl_2 or THF (~175 ml) and then precipitating it by slowly adding hexane (1000 ml) while stirring (yield 51g; 80% overall). It can also be recrystallized (e.g., toluene-hexane), but this reduces the yield.

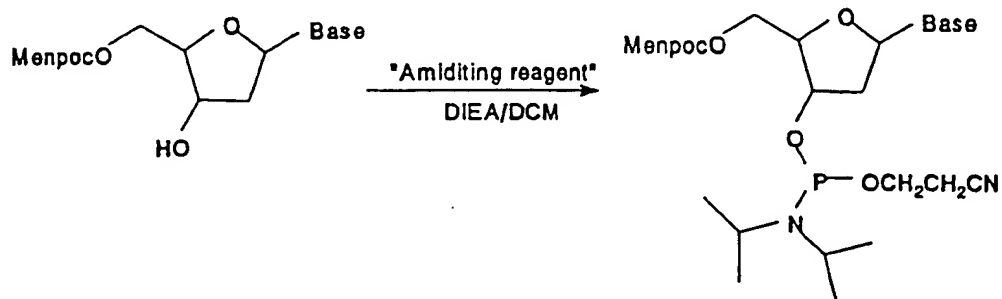
4. Synthesis of 5'-Menpoc-2'-deoxynucleoside-3'-
(N,N-diisopropyl 2-cyanoethyl phosphoramidites
(a.) 5'-MeNPOC-Nucleosides



Base= THYMIDINE (T); N-4-ISOBUTYRYL 2'-DEOXYCYTIDINE (ibu-dC);
 N-2-PHENOXYACETYL 2'DEOXYGUANOSINE (PAC-dG); and
 10 N-6-PHENOXYACETYL 2'DEOXYADENOSINE (PAC-dA)

All four of the 5'-MeNPOC nucleosides were prepared from the base-protected 2'-deoxynucleosides by the following procedure. The protected 2'-deoxynucleoside (90 mmole) was dried by
 15 co-evaporating twice with 250 mL anhydrous pyridine. The nucleoside was then dissolved in 300 mL anhydrous pyridine (or 1:1 pyridine/DMF, for the dG^{PAC} nucleoside) under argon and cooled to -2°C in an ice bath. A solution of 24.6g (90 mmole) MenPOC-Cl in 100 mL dry THF was then added with
 20 stirring over 30 minutes. The ice bath was removed, and the solution allowed to stir overnight at room temperature (TLC: 5-10% MeOH in CH₂Cl₂, two diastereomers). After evaporating the solvents under vacuum, the crude material was taken up in 250 mL ethyl acetate and extracted with saturated aqueous
 25 NaHCO₃ and brine. The organic phase was then dried over Na₂SO₄, filtered and evaporated to obtain a yellow foam. The crude products were finally purified by flash chromatography (9 x 30 cm silica gel column eluted with a stepped gradient of 2% - 6% MeOH in CH₂Cl₂). Yields of the purified diastereomeric
 30 mixtures are in the range of 65-75%.

(b.) 5'-Menpoc-2'-deoxynucleoside-3'-(N,N-diisopropyl 2-cyanoethyl phosphoramidites)



5

The four deoxynucleosides were phosphitylated using either 2-cyanoethyl- N,N- diisopropyl chlorophosphoramidite, or 2-cyanoethyl- N,N,N',N'- tetraisopropylphosphorodiamidite. The following is a typical procedure. Add 16.6g (17.4 ml; 55 mmole) of 2- cyanoethyl- N,N,N',N'- tetraisopropylphosphorodiamidite to a solution of 50 mmole 5'- MenPOC-nucleoside and 4.3g (25 mmole) diisopropylammonium tetrazolide in 250 mL dry CH₂Cl₂ under argon at ambient temperature. Continue stirring for 4-16 hours (reaction monitored by TLC: 45:45:10 hexane/CH₂Cl₂/Et₃N). Wash the organic phase with saturated aqueous NaHCO₃ and brine, then dry over Na₂SO₄, and evaporate to dryness. Purify the crude amidite by flash chromatography (9 x 25 cm silica gel column eluted with hexane/CH₂Cl₂/TEA - 45:45:10 for A, C, T; or 0:90:10 for G). The yield of purified amidite is about 90%.

B. PREPARATION OF LABELED DNA/HYBRIDIZATION TO ARRAY

25

1. PCR

PCR amplification reactions are typically conducted in a mixture composed of, per reaction: 1 μ l genomic DNA; 10 μ l each primer (10 pmol/ μ l stocks); 10 μ l 10 x PCR buffer (100 mM Tris.Cl pH8.5, 500 mM KCl, 15 mM MgCl₂); 10 μ l 2 mM dNTPs (made from 100 mM dNTP stocks); 2.5 U Taq polymerase (Perkin Elmer AmpliTaq™, 5 U/ μ l); and H₂O to 100 μ l. The cycling conditions are usually 40 cycles (94°C 45 sec, 55°C 30 sec, 72°C 60 sec) but may need to be varied considerably from

sample type to sample type. These conditions are for 0.2 mL thin wall tubes in a Perkin Elmer 9600 thermocycler. See Perkin Elmer 1992/93 catalogue for 9600 cycle time information. Target, primer length and sequence composition, among other factors, may also affect parameters.

For products in the 200 to 1000 bp size range, check 2 μ l of the reaction on a 1.5% 0.5x TBE agarose gel using an appropriate size standard (ϕ X174 cut with *Hae*III is convenient). The PCR reaction should yield several picomoles of product. It is helpful to include a negative control (i.e., 1 μ l TE instead of genomic DNA) to check for possible contamination. To avoid contamination, keep PCR products from previous experiments away from later reactions, using filter tips as appropriate. Using a set of working solutions and storing master solutions separately is helpful, so long as one does not contaminate the master stock solutions.

For simple amplifications of short fragments from genomic DNA it is, in general, unnecessary to optimize Mg^{2+} concentrations. A good procedure is the following: make a master mix minus enzyme; dispense the genomic DNA samples to individual tubes or reaction wells; add enzyme to the master mix; and mix and dispense the master solution to each well, using a new filter tip each time.

25 2. PURIFICATION

Removal of unincorporated nucleotides and primers from PCR samples can be accomplished using the Promega Magic PCR Preps DNA purification kit. One can purify the whole sample, following the instructions supplied with the kit (proceed from section IIIB, 'Sample preparation for direct purification from PCR reactions'). After elution of the PCR product in 50 μ l of TE or H_2O , one centrifuges the eluate for 20 sec at 12,000 rpm in a microfuge and carefully transfers 45 μ l to a new microfuge tube, avoiding any visible pellet. Resin is sometimes carried over during the elution step. This transfer prevents accidental contamination of the linear amplification reaction with 'Magic PCR' resin. Other methods, e.g., size exclusion chromatography, may also be used.

3. Linear amplification

In a 0.2 mL thin-wall PCR tube mix: 4 μ l purified PCR product; 2 μ l primer (10 pmol/ μ l); 4 μ l 10 x PCR buffer; 4 μ l dNTPs (2 mM dA, dC, dG, 0.1 mM dT); 4 μ l 0.1 mM dUTP; 1 μ l 1 mM fluorescein dUTP (Amersham RPN 2121); 1 U Taq polymerase (Perkin Elmer, 5 U/ μ l); and add H₂O to 40 μ l. Conduct 40 cycles (92°C 30 sec, 55°C 30 sec, 72°C 90 sec) of PCR. These conditions have been used to amplify a 300 nucleotide mitochondrial DNA fragment but are applicable to other fragments. Even in the absence of a visible product band on an agarose gel, there should still be enough product to give an easily detectable hybridization signal. If one is not treating the DNA with uracil DNA glycosylase (see Section 4), dUTP can be omitted from the reaction.

4. Fragmentation

Purify the linear amplification product using the Promega Magic PCR Preps DNA purification kit, as per Section 2 above. In a 0.2 mL thin-wall PCR tube mix: 40 μ l purified labeled DNA; 4 μ l 10 x PCR buffer; and 0.5 μ l uracil DNA glycosylase (BRL 1U/ μ l). Incubate the mixture 15 min at 37°C, then 10 min at 97°C; store at -20°C until ready to use.

5. Hybridization, Scanning & Stripping

A blank scan of the slide in hybridization buffer only is helpful to check that the slide is ready for use. The buffer is removed from the flow cell and replaced with 1 mL of (fragmented) DNA in hybridization buffer and mixed well. The scan is performed in the presence of the labeled target. Fig. 51 illustrates an illustrative detection system for scanning a DNA chip. A series of scans at 30 min intervals using a hybridization temperature of 25°C yields a very clear signal, usually in at least 30 min to two hours, but it may be desirable to hybridize longer, i.e., overnight. Using a laser power of 50 μ W and 50 μ m pixels, one should obtain maximum counts in the range of hundreds to low thousands/pixel for a new slide. When finished, the slide can be stripped using 50%

formamide. rinsing well in deionized H₂O, blowing dry, and storing at room temperature.

C. PREPARATION OF LABELED RNA/HYBRIDIZATION TO ARRAY

5 1. Tagged primers

The primers used to amplify the target nucleic acid should have promoter sequences if one desires to produce RNA from the amplified nucleic acid. Suitable promoter sequences are shown below and include:

10 (1) the T3 promoter sequence:

5'-CGGAATTAACCCTCACTAAAGG

5'-AATTAACCCTCACTAAAGGGAG;

(2) the T7 promoter sequence:

5' TAATACGACTCACTATAGGGAG;

15 and (3) the SP6 promoter sequence:

5' ATTTAGGTGACACTATAGAA.

The desired promoter sequence is added to the 5' end of the PCR primer. It is convenient to add a different promoter to
20 each primer of a PCR primer pair so that either strand may be transcribed from a single PCR product.

Synthesize PCR primers so as to leave the DMT group on. DMT-on purification is unnecessary for PCR but appears to be important for transcription. Add 25 µl 0.5M NaOH to
25 collection vial prior to collection of oligonucleotide to keep the DMT group on. Deprotect using standard chemistry -- 55°C overnight is convenient.

HPLC purification is accomplished by drying down the oligonucleotides, resuspending in 1 mL 0.1 M TEAA (dilute 2.0
30 M stock in deionized water, filter through 0.2 micron filter) and filter through 0.2 micron filter. Load 0.5 mL on reverse phase HPLC (column can be a Hamilton PRP-1 semi-prep, #79426).
The gradient is 0 -> 50% CH₃CN over 25 min (program 0.2
µmol.prep.0-50, 25 min). Pool the desired fractions, dry down,
35 resuspend in 200 µl 80% HAc. 30 min RT. Add 200 µl EtOH; dry down. Resuspend in 200 µl H₂O, plus 20 µl NaAc pH5.5, 600 µl EtOH. Leave 10 min on ice; centrifuge 12,000 rpm for 10 min in microfuge. Pour off supernatant. Rinse pellet with 1 mL

EtOH, dry, resuspend in 200 μ l H₂O. Dry, resuspend in 200 μ l TE. Measure A₂₆₀, prepare a 10 pmol/ μ l solution in TE (10 mM Tris.Cl pH 8.0, 0.1 mM EDTA). Following HPLC purification of a 42 mer, a yield in the vicinity of 15 nmol from a 0.2 μ mol scale synthesis is typical.

2. Genomic DNA Preparation

Add 500 μ l (10 mM Tris.Cl pH8.0, 10 mM EDTA, 100 mM NaCl, 2% (w/v) SDS, 40 mM DTT, filter sterilized) to the sample. Add 1.25 μ l 20 mg/ml proteinase K (Boehringer) Incubate at 55°C for 2 hours, vortexing once or twice. Perform 2x 0.5 mL 1:1 phenol:CHCl₃ extractions. After each extraction, centrifuge 12,000 rpm 5 min in a microfuge and recover 0.4 mL supernatant. Add 35 μ l NaAc pH5.2 plus 1 mL EtOH. Place sample on ice 45 min; then centrifuge 12,000 rpm 30 min, rinse, air dry 30 min, and resuspend in 100 μ l TE.

3. PCR

PCR is performed in a mixture containing, per reaction: 1 μ l genomic DNA; 4 μ l each primer (10 pmol/ μ l stocks); 4 μ l 10 x PCR buffer (100 mM Tris.Cl pH8.5, 500 mM KCl, 15 mM MgCl₂); 4 μ l 2 mM dNTPs (made from 100 mM dNTP stocks); 1 U Taq polymerase (Perkin Elmer, 5 U/ μ l); H₂O to 40 μ l. About 40 cycles (94°C 30 sec, 55°C 30 sec, 72°C 30 sec) are performed, but cycling conditions may need to be varied. These conditions are for 0.2 mL thin wall tubes in Perkin Elmer 9600. For products in the 200 to 1000 bp size range, check 2 μ l of the reaction on a 1.5% 0.5xTBE agarose gel using an appropriate size standard. For larger or smaller volumes (20 - 100 μ l), one can use the same amount of genomic DNA but adjust the other ingredients accordingly.

4. In vitro transcription

Mix: 3 μ l PCR product; 4 μ l 5x buffer; 2 μ l DTT; 2.4 μ l 10 mM rNTPs (100 mM solutions from Pharmacia); 0.48 μ l 10 mM fluorescein-UTP (Fluorescein-12-UTP, 10 mM solution, from Boehringer Mannheim); 0.5 μ l RNA polymerase (Promega T3 or T7 RNA polymerase); and add H₂O to 20 μ l. Incubate at 37°C for 3

h. Check 2 μ l of the reaction on a 1.5% 0.5xTBE agarose gel using a size standard. 5x buffer is 200 mM Tris pH 7.5, 30 mM $MgCl_2$, 10 mM spermidine, 50 mM NaCl, and 100 mM DTT (supplied with enzyme). The PCR product needs no purification and can be added directly to the transcription mixture. A 20 μ l reaction is suggested for an initial test experiment and hybridization; a 100 μ l reaction is considered "preparative" scale (the reaction can be scaled up to obtain more target). The amount of PCR product to add is variable; typically a PCR reaction will yield several picomoles of DNA. If the PCR reaction does not produce that much target, then one should increase the amount of DNA added to the transcription reaction (as well as optimize the PCR). The ratio of fluorescein-UTP to UTP suggested above is 1:5, but ratios from 1:3 to 1:10 - all work well. One can also label with biotin-UTP and detect with streptavidin-FITC to obtain similar results as with fluorescein-UTP detection.

For nondenaturing agarose gel electrophoresis of RNA, note that the RNA band will normally migrate somewhat faster than the DNA template band, although sometimes the two bands will comigrate. The temperature of the gel can effect the migration of the RNA band. The RNA produced from *in vitro* transcription is quite stable and can be stored for months (at least) at $-20^{\circ}C$ without any evidence of degradation. It can be stored in unsterilized 6xSSPE 0.1% triton X-100 at $-20^{\circ}C$ for days (at least) and reused twice (at least) for hybridization, without taking any special precautions in preparation or during use. RNase contamination should of course be avoided. When extracting RNA from cells, it is preferable to work very rapidly and to use strongly denaturing conditions. Avoid using glassware previously contaminated with RNases. Use of new disposable plasticware (not necessarily sterilized) is preferred, as new plastic tubes, tips, etc., are essentially RNase free. Treatment with DEPC or autoclaving is typically not necessary.

5. Fragmentation

Heat transcription mixture at 94 degrees for forty min.
The extent of fragmentation is controlled by varying Mg^{2+}
concentration (30 mM is typical), temperature, and duration of
5 heating.

6. Hybridization, Scanning, & Stripping

A blank scan of the slide in hybridization buffer only is
helpful to check that the slide is ready for use. The buffer
is removed from the flow cell and replaced with 1 mL of
10 (hydrolysed) RNA in hybridization buffer and mixed well.
Incubate for 15 - 30 min at 18°C. Remove the hybridization
solution, which can be saved for subsequent experiments.
Rinse the flow cell 4 - 5 times with fresh changes of 6 x SSPE
/ 0.1% Triton X-100, equilibrated to 18°C. The rinses can be
15 performed rapidly, but it is important to empty the flow cell
before each new rinse and to mix the liquid in the cell
thoroughly. A series of scans at 30 min intervals using a
hybridization temperature of 25°C yields a very clear signal,
usually in at least 30 min to two hours, but it may be
20 desirable to hybridize longer, i.e., overnight. Using a laser
power of 50 μ W and 50 μ m pixels, one should obtain maximum
counts in the range of hundreds to low thousands/pixel for a
new slide. When finished, the slide can be stripped using
warm water.

25 These conditions are illustrative and assume a probe
length of ~15 nucleotides. The stripping conditions suggested
are fairly severe, but some signal may remain on the slide if
the washing is not stringent. Nevertheless, the counts
remaining after the wash should be very low in comparison to
30 the signal in presence of target RNA. In some cases, much
gentler stripping conditions are effective. The lower the
hybridization temperature and the longer the duration of
hybridization, the more difficult it is to strip the slide.
Longer targets may be more difficult to strip than shorter
35 targets.

7. Amplification of Signal

A variety of methods can be used to enhance detection of
labelled targets bound to a probe on the array. In one

embodiment, the protein MutS (from *E. coli*) or equivalent proteins such as yeast MSH1, MSH2, and MSH3; mouse Rep-3, and *Streptococcus* Hex-A, is used in conjunction with target hybridization to detect probe-target complex that contain
5 mismatched base pairs. The protein, labeled directly or indirectly, can be added to the chip during or after hybridization of target nucleic acid, and differentially binds to homo- and heteroduplex nucleic acid. A wide variety of dyes and other labels can be used for similar purposes. For
10 instance, the dye YOYO-1 is known to bind preferentially to nucleic acids containing sequences comprising runs of 3 or more G residues.

8. Detection of Repeat Sequences

15 In some circumstances, i.e., target nucleic acids with repeated sequences or with high G/C content, very long probes are sometimes required for optimal detection. In one embodiment for detecting specific sequences in a target nucleic acid with a DNA chip, repeat sequences are detected as
20 follows. The chip comprises probes of length sufficient to extend into the repeat region varying distances from each end. The sample, prior to hybridization, is treated with a labelled oligonucleotide that is complementary to a repeat region but shorter than the full length of the repeat. The target
25 nucleic is labelled with a second, distinct label. After hybridization, the chip is scanned for probes that have bound both the labelled target and the labelled oligonucleotide probe; the presence of such bound probes shows that at least two repeat sequences are present.

30 While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made
35 without departing from the true scope of the invention. All publications and patent documents cited in this application are incorporated by reference in their entirety for all

purposes to the same extent as if each individual publication or patent document were so individually denoted.

WHAT IS CLAIMED IS:

General tiling claims

1 1. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least two sets of
3 oligonucleotide probes,

4 (1) a first probe set comprising a plurality of
5 probes, each probe comprising a segment of at least three
6 nucleotides exactly complementary to a subsequence of the
7 reference sequence, the segment including at least one
8 interrogation position complementary to a corresponding
9 nucleotide in the reference sequence,

10 (2) a second probe set comprising a corresponding
11 probe for each probe in the first probe set, the corresponding
12 probe in the second probe set being identical to a sequence
13 comprising the corresponding probe from the first probe set or
14 a subsequence of at least three nucleotides thereof that
15 includes the at least one interrogation position, except that
16 the at least one interrogation position is occupied by a
17 different nucleotide in each of the two corresponding probes
18 from the first and second probe sets;

19 wherein the probes in the first probe set have at least
20 two interrogation positions respectively corresponding to each
21 of two contiguous nucleotides in the reference sequence.

1 2. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least four sets of
3 oligonucleotide probes,

4 (1) a first probe set comprising a plurality of
5 probes, each probe comprising a segment of at least three
6 nucleotides exactly complementary to a subsequence of the
7 reference sequence, the segment including at least one
8 interrogation position complementary to a corresponding
9 nucleotide in the reference sequence,

10 (2) second, third and fourth probe sets, each
11 comprising a corresponding probe for each probe in the first
12 probe set, the probes in the second, third and fourth probe
13 sets being identical to a sequence comprising the
14 corresponding probe from the first probe set or a subsequence

15 of at least three nucleotides thereof that includes the at
16 least one interrogation position, except that the at least one
17 interrogation position is occupied by a different nucleotide
18 in each of the four corresponding probes from the four probe
19 sets.

1 3. The oligonucleotide array of claim 2, further
2 comprising a fifth probe set comprising a corresponding probe
3 for each probe in the first probe set, the corresponding probe
4 from the fifth probe set being identical to a sequence
5 comprising the corresponding probe from the first probe set or
6 a subsequence of at least three nucleotides thereof that
7 includes the at least one interrogation position, except that
8 the at least one interrogation position is deleted in the
9 corresponding probe from the fifth probe set.

1 4. The oligonucleotide array of claim 2, further
2 comprising a sixth probe set comprising a corresponding probe
3 for each probe in the first probe set, the corresponding probe
4 from the sixth probe set being identical to a sequence
5 comprising the corresponding probe from the first probe set or
6 a subsequence of at least three nucleotides thereof that
7 includes the at least one interrogation position, except that
8 an additional nucleotide is inserted adjacent to the at least
9 one interrogation position in the corresponding probe from the
10 first probe set.

1 5. The array of claim 2, wherein the first probe set has
2 at least three interrogation positions respectively
3 corresponding to each of three contiguous nucleotides in a
4 reference sequence.

1 6. The array of claim 2, wherein the first probe set has
2 at least 50 interrogation positions respectively corresponding
3 to each of 50 contiguous nucleotides in a reference sequence.

1 7. The array of claim 1 or 2, wherein the first probe
2 set has at least 100 interrogation positions respectively

3 corresponding to each of 100 contiguous nucleotides in a
4 reference sequence.

1 8. The oligonucleotide array of claim 1 or 2, wherein
2 the first probe set has an interrogation position
3 corresponding to each of at least 30% of the nucleotides in a
4 reference sequence and the reference sequence comprises at
5 least 100 nucleotides.

1 9. The oligonucleotide array of claim 8, wherein the
2 first probe set comprises probes which completely span the
3 reference sequence, which probes relative to the reference
4 sequence, overlap one another in sequence.

1 10. The oligonucleotide array of claim 9, wherein the
2 first probe set has an interrogation position corresponding to
3 each of the nucleotides in the reference sequence.

1 11. The oligonucleotide array of claim 10, wherein the
2 probes are oligodeoxyribonucleotides.

1 12. The oligonucleotide array of claim 1 or 2, wherein
2 the array comprises between 100 and 10,000 probes.

1 13. The oligonucleotide array of claim 1 or 2, wherein
2 the array comprises between 10,000 and 100,000 probes.

1 14. The oligonucleotide array of claim 1 or 2, wherein
2 the array comprises between 100,000 and 10,000,000 probes.

1 15. The oligonucleotide array of claim 1 or 2, wherein
2 the probes are linked to the support via a spacer.

1 16. The oligonucleotide array of claim 1 or 2, wherein
2 the segment in each probe of the first probe set that is
3 exactly complementary to the subsequence of the reference
4 sequence is 9-21 nucleotides.

1 17. The oligonucleotide array of claim 16, wherein the
2 segment is n nucleotides long, and the subsequence is at least
3 n-2 nucleotides long.

1 18. The oligonucleotide array of claim 1 or 2, wherein
2 each probe of the first probe set consists of the segment that
3 is exactly complementary to the subsequence of the reference
4 sequence.

1 19. The oligonucleotide array of claim 1 or 2, wherein
2 the probes in the second, third and fourth probe sets are
3 identical to the corresponding probe from the first probe set
4 except that the at least one interrogation position is
5 occupied by a different nucleotide in each of the four
6 corresponding probes from the four probe sets.

1 20. The array of claim 2, further comprising fifth,
2 sixth and seventh probe sets, wherein:
3 the segment of each probe in the first set
4 includes at least two interrogation positions each
5 corresponding to a nucleotide in the reference sequence,
6 the second, third and fourth probe sets, each
7 comprise a corresponding probe for each probe in the first
8 probe set, the corresponding probes in the second, third and
9 fourth probe sets being identical to a sequence comprising the
10 corresponding probe from the first probe set or a subsequence
11 of at least three nucleotides thereof that includes a first
12 interrogation position except that the first interrogation
13 position is occupied by a different nucleotide in each of the
14 four corresponding probes from the four probe sets;
15 the fifth, sixth and seventh probe sets, each
16 comprising a corresponding probe for each probe in the first
17 probe set, the probes in the fifth, sixth and seventh probe
18 sets being identical to a sequence comprising the
19 corresponding probe from the first probe set or a subsequence
20 of at least three nucleotides thereof that includes a second
21 interrogation position, except that the second interrogation

22 position is occupied by a different nucleotide in each of the
23 four corresponding probes from the four probe sets.

1 21. The array of claim 2, wherein each probe in the
2 first probe set further comprises a second segment of at least
3 three nucleotides exactly complementary to a second
4 subsequence of the reference sequence, and the probes from the
5 second, third and fourth probe sets comprise the corresponding
6 probe from the first probe set or a subsequence thereof
7 comprising the first and second segments except in the at
8 least one interrogation position.

1 22. The array of claim 2, further comprising:
2 a fifth probe set comprising at least one probe
3 comprising a segment of at least seven nucleotides exactly
4 complementary to a subsequence of the reference sequence
5 except at one or two positions, the segment including at least
6 one interrogation position corresponding to a nucleotide in
7 the reference sequence not at the one or two positions;
8 sixth, seventh and eighth probe sets, each comprising a
9 probe for each probe in the fifth probe set, the corresponding
10 probes from the sixth, seventh & eighth probe sets being
11 identical to a sequence comprising the corresponding probe
12 from the fifth probe set or a subsequence of at least nine
13 nucleotides thereof including the at least one interrogation
14 position and the one or two positions, except in the at least
15 one interrogation position, which is occupied by a different
16 nucleotide in each of the four probes.

1 23. The array of claim 2, wherein the probes are
2 arranged on the substrate so that the first set of probes is
3 arranged in a row across the substrate in an order reflecting
4 the overlap between the probes and the reference sequence, and
5 the additional sets of probes are arranged in columns relative
6 to the probes in said first set, so that probes with the same
7 interrogation position are in the same column and so that each
8 column comprises at least 4 probes.

1 24. The array of Claim 2, wherein said probes are 12 to
2 17 nucleotides in length.

1 25. The array of Claim 2, wherein said probes are 15
2 nucleotides in length and attached by a covalent linkage to a
3 site on a 3'-end of said probes, and said interrogation
4 position is located at position 7, relative to the 3'-end of
5 said probes.

1 26. The array of claim 2, further comprises fifth,
2 sixth, seventh and eighth probe sets,

3 (1) a fifth probe set comprising a plurality of
4 probes, each probe comprising a segment of at least three
5 nucleotides exactly complementary to a subsequence of a second
6 reference sequence, the segment including at least one
7 interrogation position complementary to a corresponding
8 nucleotide in the reference sequence,

9 (2) the sixth, seventh, and eighth probe sets, each
10 comprising a corresponding probe for each probe in the fifth
11 probe set, the probes in the sixth, seventh and eighth probe
12 sets being identical to a sequence comprising the
13 corresponding probe from the fifth probe set or a subsequence
14 of at least three nucleotides thereof that includes the at
15 least one interrogation position, except that the at least one
16 interrogation position is occupied by a different nucleotide
17 in each of the four corresponding probes from the fifth,
18 sixth, seventh and eighth probe sets.

1 27. The array of claim 22, wherein the first, second,
2 third and fourth probe sets have probes of a first length and
3 the fifth, sixth, seventh and eighth probe sets have probes of
4 a second length different from the first length.

Tiling for wildtype and mutant reference sequences

1 28. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least one pair of first
3 and second probe groups, each group comprising a first and
4 second sets of oligonucleotide probes as defined by claim 1;

5 wherein each probe in the first probe set from the
6 first group is exactly complementary to a subsequence of a
7 first reference sequence and each probe in the first probe set
8 from the second group is exactly complementary to a
9 subsequence from a second reference sequence.

1 29. The array of claim 28, wherein the second reference
2 sequence is a mutated form of the first reference sequence.

1 30. The array of claim 28, wherein each group further
2 comprises third and fourth probe sets, each comprising a
3 corresponding probe for each probe in the first probe set, the
4 probes in the second, third and fourth probe sets being
5 identical to a sequence comprising the corresponding probe
6 from the first probe set or a subsequence of at least three
7 nucleotides thereof that includes the interrogation position,
8 except that the interrogation position is occupied by a
9 different nucleotide in each of the four corresponding probes
10 from the four probe sets.

1 31. The array of claim 30 that comprises at least five
2 pairs of first and second probe groups, wherein the probes in
3 the first probe sets from the first groups of the five pairs
4 are exactly complementary to subsequences from five different
5 respective first reference sequences.

1 32. The array of claim 30 that comprises at least forty
2 pairs of first and second probe groups, wherein the probes in
3 the first probe sets from the first groups of the forty pairs
4 are exactly complementary to subsequences from forty
5 respective first reference sequences.

Block tiling

1 33. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least a group of probes
3 comprising:

4 a wildtype probe comprising a segment of at least three
5 nucleotides exactly complementary to a subsequence of a

6 reference sequence, the segment having at least first and
7 second interrogation positions corresponding to first and
8 second nucleotides in the reference sequence,

9 a first set of three mutant probes, each identical to a
10 sequence comprising the wildtype probe or a subsequence of at
11 least three nucleotides thereof including the first and second
12 interrogation positions, except in the first interrogation
13 position, which is occupied by a different nucleotide in each
14 of the three mutant probes and the wildtype probe;

15 a second set of three mutant probes, each identical to a
16 sequence comprising the wildtype probe or a subsequence of at
17 least three nucleotides thereof including the first and second
18 interrogation positions, except in the second interrogation
19 position, which is occupied by a different nucleotide in each
20 of the three mutant probes and the wildtype probe.

1 34. The array of claim 33, wherein the segment of the
2 wildtype probe comprises 3-20 interrogation positions
3 corresponding to 3-20 respective nucleotides in the reference
4 sequence, and the array comprises 3-20 respective sets of
5 three mutant probes, each of the three probes identical to a
6 sequence comprising the wildtype probe or a subsequence
7 thereof including the 3-20 interrogation positions, except
8 that one of the 3-20 interrogation positions is occupied by a
9 different nucleotide in each of the three mutant probes and
10 the wildtype probes, the one of the 3-20 interrogation
11 positions being different in each of the 3-20 respective sets
12 of three mutant probes.

1 35. An array of probes immobilized to a solid support
2 comprising two groups of probes, each group as defined by
3 claim 33, a first group comprising a wildtype probe comprising
4 a segment exactly complementary to a subsequence of a first
5 reference sequence and a second group comprising a wildtype
6 probe comprising a segment exactly complementary to a
7 subsequence of a second reference sequence.

1 36. The array of claim 35, comprising at least 10-100
2 groups of probes, each comprising a wildtype probe comprising
3 a segment exactly complementary to a subsequence of at least
4 10-100 respective reference sequences.

Pooled probes

1 37. A method of comparing a target sequence with a
2 reference sequence, the method comprising:
3 identifying variants of a reference sequence differing
4 from the reference sequence in at least one nucleotide;
5 assigning each variant a designation,
6 providing an array of pools of probes, each pool
7 occupying a separate cell of the array, wherein each pool
8 comprises a probe comprising a segment exactly complementary
9 to each variant sequence assigned a particular designation,
10 contacting the array with a target sequence comprising a
11 variant of the reference sequence;
12 determining the relative hybridization intensities of the
13 pools in the array to the target sequence;
14 determining the target sequence from the relative
15 hybridization intensities of the pools.

1 38. The method of claim 37, wherein the variants are
2 assigned numbers according to an error code.

1 39. The method of claim 37, wherein each variant is
2 assigned a designation having at least one digit and at least
3 one value for the digit, and each pool comprise a probe
4 comprising a segment exactly complementary to each variant
5 sequence assigned a particular value in a particular digit.

1 40. The method of claim 39, wherein the variants are
2 assigned successive numbers in a numbering system of base m
3 having n digits, and the array comprises n x (m-1) pools of
4 probes.

1 41. The method of claim 40, wherein each pool further
2 comprises a probe comprising a segment exactly complementary
3 to the reference sequence.

4 **Trellis tiling**

5 42. A pooled probe comprising a segment exactly
6 complementary to a subsequence of a reference sequence except
7 at a first interrogation position occupied by a pooled
8 nucleotide N, a second interrogation position occupied by a
9 pooled nucleotide selected from the group of three consisting
10 of (1) M or K, (2) R or Y and (3) S or W, and a third
11 interrogation position occupied by a second pooled nucleotide
12 selected from the group, wherein the pooled nucleotide
13 occupying the second interrogation position comprises a
14 nucleotide complementary to a corresponding nucleotide from
15 the reference sequence when the second pooled probe and
16 reference sequence are maximally aligned, and the pooled
17 nucleotide occupying the third interrogation position
18 comprises a nucleotide complementary to a corresponding
nucleotide from the reference sequence when the third pooled
probe and the reference sequence are maximally aligned,
wherein N is A, C, G or T(U), K is G or T(U), M is A or C, R
is A or G, Y is C or T(U), W is A or T(U) and S is G or C.

1 43. An array of oligonucleotide probes immobilized on
2 solid support, the array comprising:
3 first, second and third cells respectively occupied by
4 first, second and third pooled probes, each pooled probe
5 comprising a segment exactly complementary to a subsequence of
6 a reference sequence except at a first interrogation position
7 occupied by a pooled nucleotide N, a second interrogation
8 position occupied by a pooled nucleotide selected from the
9 group of three consisting of (1) M or K, (2) R or Y and (3) S
10 or W, and a third interrogation position occupied by a second
11 pooled nucleotide selected from the group, wherein the pooled
12 nucleotide occupying the second interrogation position
13 comprises a nucleotide complementary to a corresponding
14 nucleotide from the reference sequence when the pooled probe

15 and the reference sequence are maximally aligned, and the
16 pooled nucleotide occupying the third interrogation position
17 comprises a nucleotide complementary to a corresponding
18 nucleotide from the reference sequence when the pooled probe
19 and the reference sequence are maximally aligned;
20 provided that one of the three interrogation
21 positions in the each of the three pooled probes is aligned
22 with the same corresponding nucleotide in the reference
23 sequence, this interrogation position being occupied by an N
24 in one of the pooled probes, and a different pooled nucleotide
25 in each of the other two pooled probes,
26 wherein N is A, C, G or T(U), K is G or T(U), M is A
27 or C, R is A or G, Y is C or T(U), W is A or T(U) and S is G
28 or C.

1 44. The array of claim 43 further comprising:
2 fourth and fifth cells respectively occupied by fourth
3 and fifth pooled probes, each pooled probe as defined by
4 claim 43,
5 wherein one of the three interrogation position in the
6 second, third and fourth pooled probes is aligned with the
7 same corresponding nucleotide in the reference sequence, this
8 interrogation position being occupied by an N in one of the
9 pooled probes, and a different pooled nucleotide in each of
10 the other two pooled probes,
11 wherein one of the three interrogation position in the
12 third, fourth and fifth pooled probes is aligned with the same
13 corresponding nucleotide in the reference sequence, this
14 interrogation position being occupied by an N in one of the
15 pooled probes, and a different pooled nucleotide in each of
16 the other two pooled probes.

1 45. The array of claim 44, wherein the pooled probes are
2 identical except at the interrogation positions.

1 46. The array of claim 44, wherein the first, second,
2 third, fourth and fifth pooled probes are exactly
3 complementary to five respective subsequences of the reference

4 sequences that from each other by increments of one
5 nucleotide.

Bridge tiling

: 1 47. An array of oligonucleotide probes immobilized on a
- 2 solid support, the array comprising at least four probes:
• 3 a first probe comprising first and second segments, each
: 4 of at least three nucleotides and exactly complementary to
5 first and second subsequences of a reference sequences, the
6 segments including at least one interrogation position
7 corresponding to a nucleotide in the reference sequence,
8 wherein either (1) the first and second subsequences are
9 noncontiguous, or (2) the first and second subsequences are
10 contiguous and the first and second segments are inverted
11 relative to the complement of the first and second
12 subsequences in the reference sequence;
13 second, third and fourth probes, identical to a sequence
14 comprising the first probe or a subsequence thereof comprising
15 at least three nucleotides from each of the first and second
16 segments, except in the at least one interrogation position,
17 which differs in each of the probes.

1 48. The array of claim 47, wherein the first and second
2 subsequences are separated by one or two nucleotides in the
3 reference sequence.

Two interrogation positions (no wildtype)

1 49. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least a set of four
3 probes, each of the probes comprising a segment of at least 7
4 nucleotides that is exactly complementary to a subsequence
5 from a reference sequence, except that the segment may or may
: 6 not be exactly complementary at two interrogation positions,
7 wherein:
: 8 the first interrogation position is occupied by a
9 different nucleotide in each of the four probes,
10 the second interrogation position is occupied by a
11 different nucleotide in each of the four probes,

12 in first and second probes, the segment is exactly
13 complementary to the subsequence, except at not more than one
14 of the interrogation positions, and

15 in third and fourth probes, the segment is exactly
16 complementary to the subsequence, except at both of the
17 interrogation positions.

1 50. An array of probes immobilized to a support, the
2 array comprising at least 100 sets of 4 probes, each set as
3 defined by claim 49, the probes from the at least 100 sets
4 comprising at least 100 respective segments, the segments
5 having at least 100 respective first and second interrogation
6 positions.

Helper mutations

1 51. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising a set of probes
3 comprising:

4 a first probe comprising a segment of at least 7
5 nucleotides exactly complementary to a subsequence of a
6 reference sequence except at one or two positions, the segment
7 including an interrogation position not at the one or two
8 positions;

9 second, third and fourth mutant probes, each identical to
10 a sequence comprising the wildtype probe or a subsequence
11 thereof including the interrogation position and the one or
12 two positions, except in the interrogation position, which is
13 occupied by a different nucleotide in each of the four probes.

Omission of Perfectly Matched Probe

1 52. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least two sets of
3 oligonucleotide probes,

4 (1) a first probe set comprising a plurality of
5 probes, each probe comprising a segment exactly complementary
6 to a subsequence of at least 3 nucleotides of a reference
7 sequence except at an interrogation position,

8 (2) a second probe set comprising a corresponding
9 probe for each probe in the first probe set, the corresponding
10 probe in the second probe set being identical to a sequence
11 comprising the corresponding probe from the first probe set or
12 a subsequence of at least three nucleotides thereof that
13 includes the interrogation position, except that the
14 interrogation position is occupied by a different nucleotide
15 in each of the two corresponding probes and the complement to
16 the reference sequence,

17 wherein the probes in the first probe set have at
18 least three interrogation positions respectively corresponding
19 to each of three contiguous nucleotides in the reference
20 sequence.

Methods

1 53. A method of comparing a target nucleic acid with a
2 reference sequence comprising a predetermined sequence of
3 nucleotides, the method comprising:

4 (a) hybridizing the target nucleic acid to an array
5 of oligonucleotide probes immobilized on a solid support, the
6 array comprising:

7 (1) a first probe set comprising a plurality of
8 probes, each probe comprising a segment of at least three
9 nucleotides exactly complementary to a subsequence of the
10 reference sequence, the segment including at least one
11 interrogation position complementary to a corresponding
12 nucleotide in the reference sequence,

13 (2) a second probe set comprising a corresponding
14 probe for each probe in the first probe set, the corresponding
15 probe in the second probe set being identical to a sequence
16 comprising the corresponding probe from the first probe set or
17 a subsequence of at least three nucleotides thereof that
18 includes the at least one interrogation position, except that
19 the at least one interrogation position is occupied by a
20 different nucleotide in each of the two corresponding probes
21 from the first and second probe sets;

22 wherein, the probes in the first probe set have at
23 least three interrogation positions respectively corresponding

24 to each of at least three nucleotides in the reference
25 sequence, and
26 (b) determining which probes, relative to one
27 another, in the array bind specifically to the target nucleic
28 acid, the relative specific binding of the probes indicating
29 whether the target sequence is the same or different from the
30 reference sequence.

1 54. The method of claim 53, wherein the array further
2 comprises third and fourth probe sets, each comprising a
3 corresponding probe for each probe in the first probe set, the
4 probes in the second, third and fourth probe sets being
5 identical to a sequence comprising the corresponding probe
6 from the first probe set or a subsequence of at least three
7 nucleotides thereof that includes the at least one
8 interrogation position, except that the at least one
9 interrogation position is occupied by a different nucleotide
10 in each of the four corresponding probes from the four probe
11 sets.

1 55. The method of claim 54, wherein the target sequence
2 has a substituted nucleotide relative to the reference
3 sequence in at least one undetermined position, and the
4 relative specific binding of the probes indicates the location
5 of the position and the nucleotide occupying the position in
6 the target sequence.

1 56. The method of claim 54, wherein:
2 the hybridizing step comprises hybridizing the
3 target nucleic acid and a second target nucleic acid to the
4 array; and
5 the determining step comprises determining which
6 probes, relative to one another, in the array bind
7 specifically to the target nucleic acid or the second target
8 nucleic acid, the relative specific binding of the probes
9 indicating whether the target sequence is the same or
10 different from the reference sequence and whether the second

11 target sequence is the same or different from the reference
12 sequence.

1 57. The method of claim 56, wherein the target sequence
2 has a label and the second target sequence has a second label
3 different from the label.

1 58. The method of claim 56, wherein undetermined first
2 and second proportions of the first and second target
3 sequences are hybridized to the array and the specific binding
4 indicates the proportions.

1 59. The method of claim 54, further comprising:
2 (c) removing the target nucleic acid from the array;
3 (d) hybridizing a second target nucleic acid to the
4 array;
5 (e) determining which probes, relative to one another, in
6 the array bind specifically to the second target nucleic acid,
7 the relative specific binding of the probes indicating whether
8 the second target sequence is the same or different from the
9 reference sequence.

1 60. A method of comparing a target nucleic acid with a
2 reference sequence comprising a predetermined sequence of
3 nucleotides, the method comprising:
4 hybridizing the target sequence to the array of
5 claim 28;
6 determining which probes in the first group,
7 relative to one another, hybridize to the target sequence, the
8 relative specific binding of the probes indicating whether the
9 target sequence is the same or different from the first
10 reference sequence;
11 determining which probes in the second group,
12 relative to one another, hybridize to the target sequence, the
13 relative specific binding of the probes indicating whether the
14 target sequence is the same or different from the second
15 reference sequence.

1 61. The method of claim 60, wherein the hybridizing step
2 comprising hybridizing the target sequence and a second target
3 sequence to the array, and the relative specific binding of
4 the probes from the first group indicates that the target is
5 identical to the first reference sequence, and the relative
6 specific binding of the probes from the second group indicates
7 that the second target sequence is identical to the second
8 reference sequence.

1 62. The method of claim 61, wherein the first and second
2 target sequences are heterozygous alleles of a gene.

Comparative hybridization

1 63. A method of comparing a target nucleic acid with a
2 reference sequence comprising a predetermined sequence of
3 nucleotides, the method comprising:
4 (a) hybridizing the reference sequence to an array
5 of oligonucleotide probes immobilized on a solid support, the
6 array comprising;
7 (1) a first probe set comprising a plurality of
8 probes, each probe comprising a segment of at least 3
9 nucleotides exactly complementary to a subsequence of the
10 reference sequence except in at least one interrogation
11 position;
12 (2) a second probe set comprising a corresponding
13 probe for each probe in the first probe set, the corresponding
14 probe in the second probe set being identical to a sequence
15 comprising the corresponding probe from the first probe set or
16 a subsequence of at least three nucleotides thereof that
17 includes the at least one interrogation position, except that
18 the at least one interrogation position is occupied by a
19 different nucleotide in each of the two corresponding probes
20 from the first and second probe sets; and
21 (b) determining which probes, relative to one
22 another, in the array bind specifically to the reference
23 sequence;
24 (c) hybridizing a target sequence to the array;

25 (d) determining which probes, relative to one
26 another, in the array bind specifically to the target
27 sequence;
28 wherein the relative specific binding of the probes
29 to the reference and the target sequence indicates whether the
30 reference sequence is the same or different from the target
31 sequence.

1 64. The method of claim 63, wherein the reference
2 sequence has a first label and the second reference sequence
3 has a second label different from the first label, and steps
4 (a) and (c) are performed simultaneously.

HIV Chip

1 65. The array of claim 2, wherein the reference sequence
2 is from a human immunodeficiency virus.

1 66. The array of claim 65, wherein the reference
2 sequence is from a reverse transcriptase gene of the human
3 immunodeficiency virus.

1 67. The array of claim 66, wherein the reference
2 sequence is from a protease gene of the human immunodeficiency
3 virus.

1 68. The array of claim 66, wherein the reference
2 sequence is a full-length reverse transcriptase gene.

1 69. The array of claim 68 comprising at least 3200
2 oligonucleotide probes.

1 70. The array of claim 66, wherein the HIV gene is from
2 the BRU HIV strain.

1 71. The array of claim 66, wherein the HIV gene is from
2 the SF2 HIV strain.

1 72. The array of claim 28, wherein the reference
2 sequence is from the coding strand of a reverse transcriptase
3 gene of a human immunodeficiency virus and the second
4 reference sequence is from the noncoding strand of the reverse
5 transcriptase gene.

1 73. The array of claim 28, wherein the first reference
2 sequence is from a reverse transcriptase gene of a human
3 immunodeficiency virus and the second reference sequence
4 comprises a subsequence of the first reference sequence with a
5 substitution of at least one nucleotide.

1 74. The array of claim 73, wherein the substitution
2 confers drug resistance to a human immunodeficiency virus
3 comprising the second reference sequence.

1 75. The array of claim 28, wherein the first and second
2 reference sequences are from a reverse transcriptase gene from
3 first and second strains of a human immunodeficiency virus.

1 76. The array of claim 28, wherein the first reference
2 sequence is from a reverse transcriptase gene of a human
3 immunodeficiency virus and the second reference sequence is
4 from a 16S RNA, or DNA encoding the 16S RNA, from a pathogenic
5 microorganism.

1 77. The array of claim 28, wherein the first reference
2 sequence is from a reverse transcriptase gene of a human
3 immunodeficiency virus and the second reference sequence is
4 from a protease gene of the human immunodeficiency virus.

1 78. The method of claim 54, wherein the reference
2 sequence is from a human immunodeficiency virus.

1 79. The method of claim 78, wherein the reference
2 sequence is from a human immunodeficiency virus and the target
3 sequence is from a second human immunodeficiency virus.

1 80. The method of claim 79, wherein the target sequence
2 has a substituted nucleotide relative to the reference
3 sequence in at least one undetermined position, and the
4 relative specific binding of the probes indicates the location
5 of the position and the nucleotide occupying the position in
6 the target sequence.

1 81. The method of claim 80, wherein the target sequence
2 has a substituted nucleotide relative to the reference
3 sequence in at least one position, the substitution conferring
4 drug resistance to the human immunodeficiency virus, and the
5 relative specific binding of the probes reveals the
6 substitution.

1 82. The method of claim 78, wherein:
2 the hybridizing step comprises hybridizing the
3 target nucleic acid and a second target nucleic acid, the
4 second target sequence being from a reverse transcriptase gene
5 of a third human immunodeficiency virus, to the array; and
6 the determining step comprises determining which
7 probes, relative to one another, in the array bind
8 specifically to the target nucleic acid or the second target
9 nucleic acid, the relative specific binding of the probes
10 indicating whether the target sequence is the same or
11 different from the reference sequence and whether the second
12 target sequence is the same or different from the reference
13 sequence.

1 83. The method of claim 82, wherein the first target
2 sequence has a first label and the second target sequence has
3 a second label different from the first label.

1 84. The method of claim 82, wherein undetermined first
2 and second proportions of the first and second target
3 sequences are hybridized to the array and the specific binding
4 indicates the proportions.

CFTR Chip

1 85. The array of claim 2, wherein the reference sequence
2 is from a CFTR gene.

1 86. The array of claim 85, wherein the reference
2 sequence is exon 10 of a CFTR gene, and said array comprises
3 over 1000 oligonucleotide probes, 10 to 18 nucleotides in
4 length.

1 87. The array of claim 85, wherein said array comprises
2 a set of probes comprising a specific nucleotide sequence
3 selected from the group of sequences comprising:
4 3'-TTTATAXTAG;
5 3'- TTATAGXAGA;
6 3'- TATAGTXGAA;
7 3'- ATAGTAXAAA;
8 3'- TAGTAGXAAC;
9 3'- AGTAGAXACC;
10 3'- GTAGAAXCCA;
11 3'- TAGAAAXCAC; and
12 3'- AGAAACXACA; wherein each set comprises 4 probes,
13 and X is individually A, G, C, and T for each set.

1 88. The array of claim 85, wherein said group of
2 sequences comprises:
3 3'-TTTATAXTAGAAACC;
4 3'- TTATAGXAGAAACCA;
5 3'- TATAGTXGAAACCAC;
6 3'- ATAGTAXAAACCACA;
7 3'- TAGTAGXAACCACAA;
8 3'- AGTAGAXACCACAAA;
9 3'- GTAGAAXCCACAAAG;
10 3'- TAGAAAXCACAAAGG; and
11 3'- AGAAACXACAAAGGA; wherein each set comprises 4
12 probes, and X is individually A, G, C, and T for each set.

1 89. The array of claim 32, wherein the forty first
2 reference sequences are from a CFTR gene.

1 90. The array of claim 89, wherein each of the forty
2 first reference sequences includes a site of a mutation and at
3 least one adjacent nucleotide.

1 91. The array of claim 90, wherein each of the forty
2 first reference sequences comprises at least five contiguous
3 nucleotides from a CFTR gene.

1 92. The array of claim 89, wherein at least one first
2 reference sequence is a from the coding strand of the cystic
3 fibrosis gene and at least one first reference sequence is
4 from the noncoding strand of the CFTR gene.

1 93. An array of oligonucleotide probes immobilized on a
2 solid support, the array comprising at least a group of probes
3 comprising:

4 a wildtype probe exactly complementary to a subsequence
5 of a reference sequence from a cystic fibrosis gene, the
6 segment having at least five interrogation positions
7 corresponding to five contiguous nucleotides in the reference
8 sequence,

9 a first set of three mutant probes, each identical to the
10 wildtype probe, except in a first of the five interrogation
11 positions, which is occupied by a different nucleotide in each
12 of the three mutant probes and the wildtype probe;

13 a second set of three mutant probes, each identical to
14 the wildtype probe, except in a second of the five
15 interrogation positions, which is occupied by a different
16 nucleotide in each of the three mutant probes and the wildtype
17 probe;

18 a third set of three mutant probes, each identical to the
19 wildtype probe, except in a third of the five interrogation
20 positions, which is occupied by a different nucleotide in each
21 of the three mutant probes and the wildtype probe;

22 a fourth set of three mutant probes, each identical to
23 the wildtype probe, except in a fourth of the five
24 interrogation positions, which is occupied by a different

25 nucleotide in each of the three mutant probes and the wildtype
26 probe;

27 a fifth set of three mutant probes, each identical to the
28 wildtype probe, except in a fifth of the five interrogation
29 positions, which is occupied by a different nucleotide in each
30 of the three mutant probes and the wildtype probe.

1 94. The array of claim 93 comprising first and second
2 groups of probes, each group as defined by claim 93, the first
3 group comprising a wildtype probe exactly complementary to a
4 first reference sequence, and the second group comprising a
5 wildtype probe exactly complementary to a second reference
6 sequence, wherein the second reference sequence is a mutated
7 form of the first reference sequence.

1 95. The array of claim 94, wherein the first reference
2 sequence is from a CFTR gene and the second reference sequence
3 is a mutated form of the first reference sequence.

1 96. The method of claim 56, wherein the target sequence
2 and the second target sequence are from heterozygous alleles
3 of a CFTR gene.

P53 Chip

1 97. The array of claim 2, wherein the reference sequence
2 is a sequence from a p53 gene.

1 98. The array of claim 2, wherein the reference sequence
2 is from an hMLH1 gene.

1 99. The array of claim 2, wherein the reference sequence
2 is from an MSH2 gene.

1 100. The array of claim 28, wherein the reference
2 sequence is from a human P53 gene and the second reference
3 sequence is from an hMLH1 gene.

1 101. The array of claim 100, further comprising:

2 ninth, tenth, eleventh and twelfth probe sets,

3 (1) the ninth probe set comprising a plurality of
4 probes, each probe comprising a segment of at least three
5 nucleotides exactly complementary to a subsequence of a third
6 reference sequence, the segment including at least one
7 interrogation position complementary to a corresponding
8 nucleotide in the third reference sequence,

9 (2) the tenth, eleventh and twelfth probe sets,
10 each comprising a corresponding probe for each probe in the
11 ninth probe set, the probes in the tenth, eleventh and twelfth
12 probe sets being identical to a sequence comprising the
13 corresponding probe from the ninth probe set or a subsequence
14 of at least three nucleotides thereof that includes the at
15 least one interrogation position, except that the at least one
16 interrogation position is occupied by a different nucleotide
17 in each of the four corresponding probes from the ninth,
18 tenth, eleventh and twelfth probe sets.

1 102. The array of claim 97, wherein the first probe set
2 has at least 60 interrogation positions corresponding to at 60
3 contiguous nucleotides from exon 6.

1 103. The array of claim 98, wherein the reference
2 sequence is exon 5 of a p53 gene, the probes are 17
3 nucleotides long, and the first set of probes is exactly
4 complementary to the reference sequence, and the at least one
5 interrogation position is at position 7, relative to a 3'-end
6 of each probe, which 3'-end is covalently attached to the
7 substrate.

Mitochondrial Chip

1 104. The array of claim 2, wherein the reference
2 sequence is from a mitochondrial genome.

1 105. The array of claim 104, wherein said reference
2 sequence is a sequence of a D-loop region.

1 106. The array of claim 105, wherein D-loop region is
2 full-length.

1 107. The array of claim 104, wherein said reference
2 sequence is at least 90% of a full-length mitochondrial
3 genome.

1 108. The array of claim 104, wherein the reference
2 sequence is bounded by positions 16280 to 356 of the
3 mitochondrial genome.

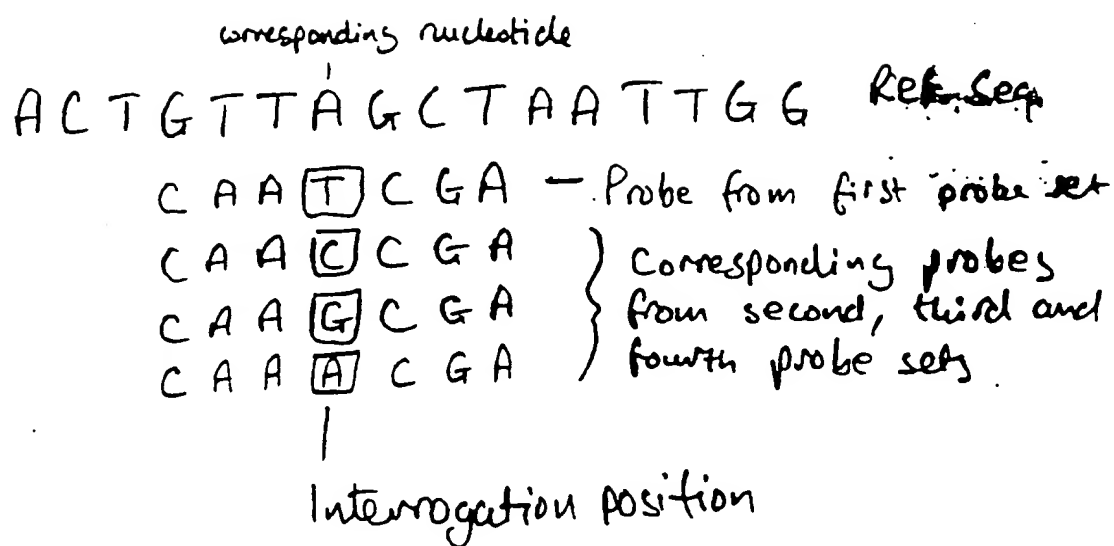


Fig. 1

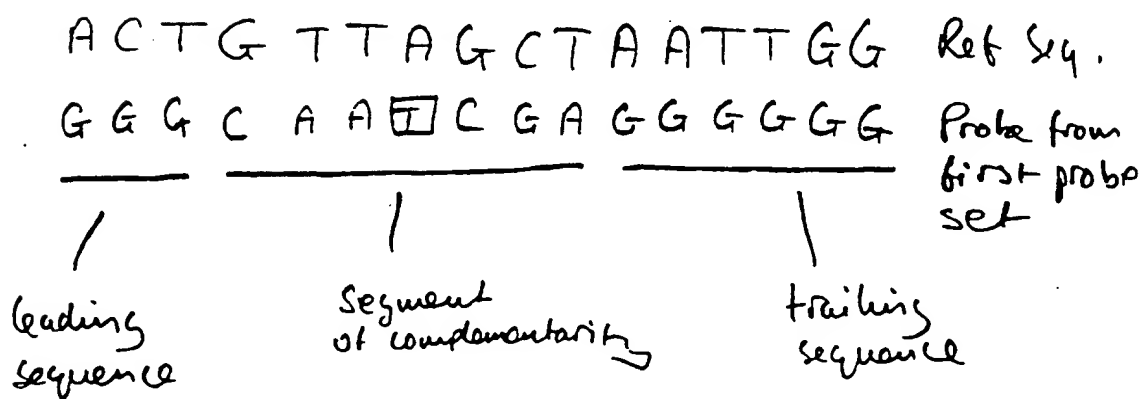
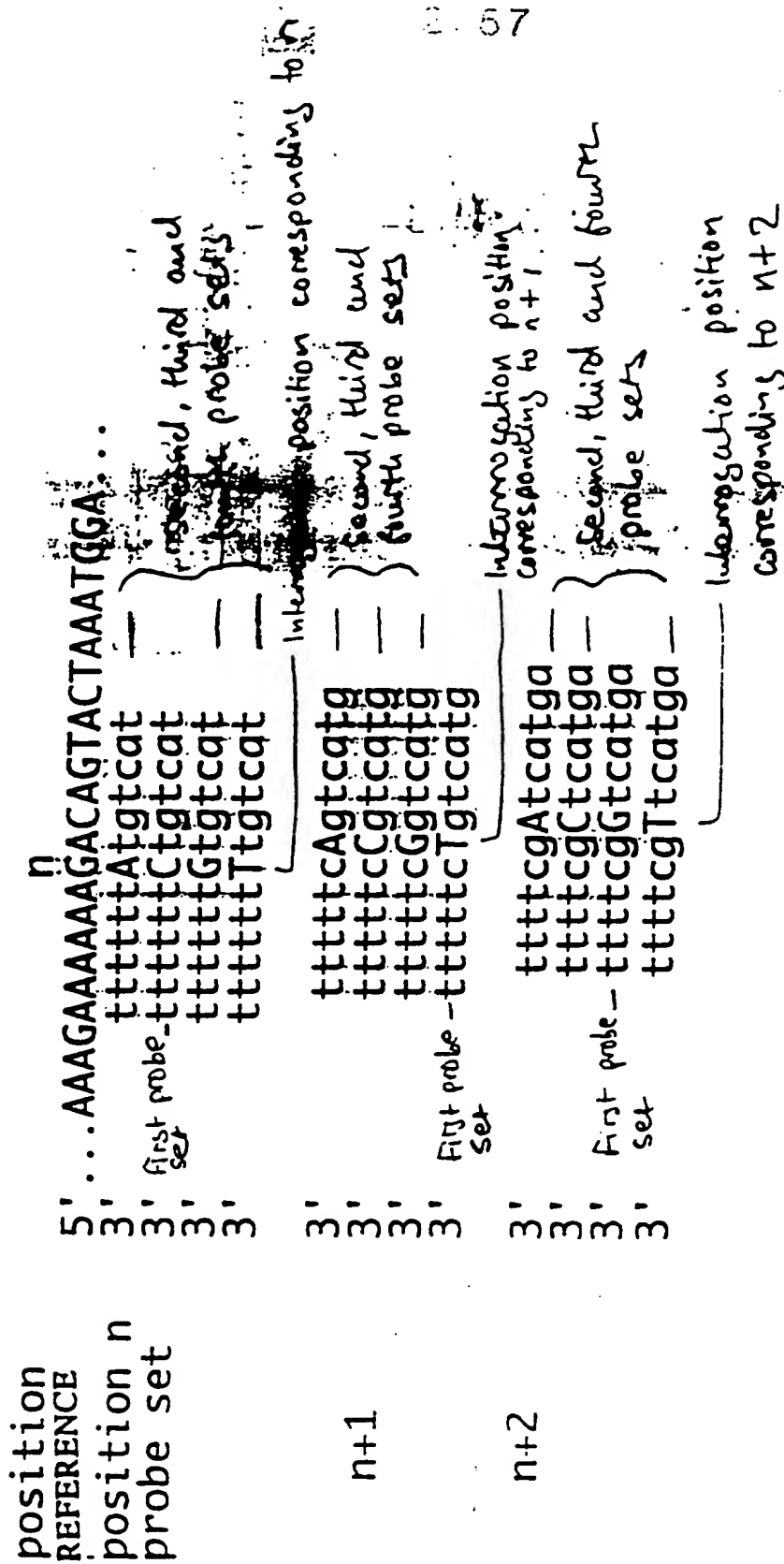


Fig. 2



3/57

$n_1 \ n_2 \ n_3 \ n_4 \ n_5$
A C T G T T A G C T A A T T G G Ref.
Seq.

A-lane T G A C G A A A A C A A C A A T A A A C
C-lane T G C C G A C A A C C A C A C T A A C G
G-lane T G G C G A G A A C G A C A G T A A G G
T-lane T G T C G A T A A C T A C A T T A A T G

| | | | |
 I_1 I_2 I_3 I_4 I_5

C-lane T G C C G A A A A C A C A C T A A G

G-lane T G G C G A G A A C G A C A G T A A G G

T-lane T G T C G A A A C A C A T A A G

$$\begin{array}{ccccc} | & | & | & | & | \\ I_1 & I_2 & I_3 & I_4 & I_5 \end{array}$$

wt. lane T G A C G A C A A C A T A A T G

Fig. 4

4/57

Fig. 5

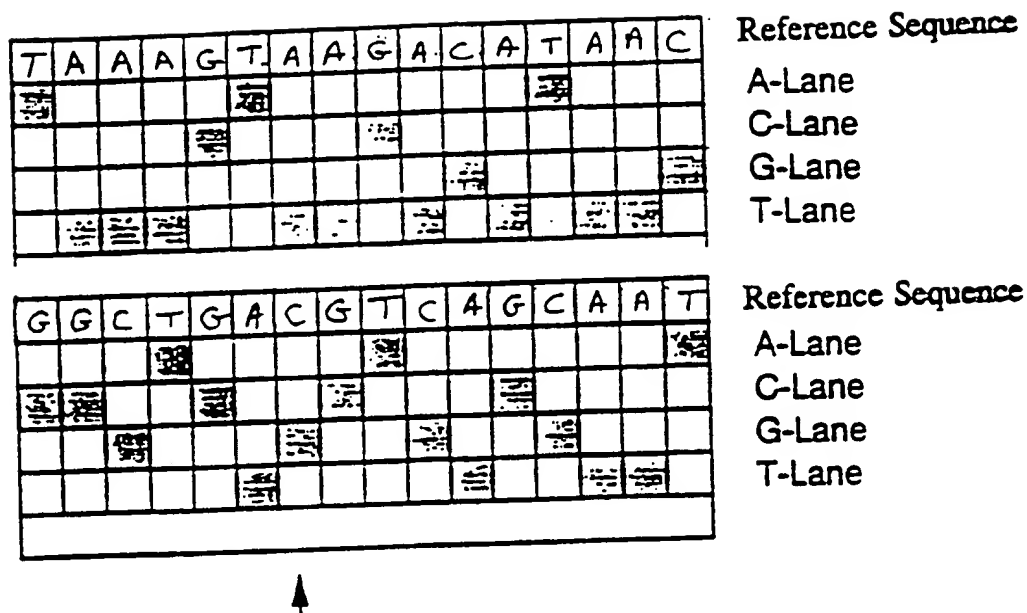


FIG. 5 : Tiled Array with Probes for the Detection of Point Mutations

3' - CCGACTACAGTCGTT
 3' - CCGACTCCAGTCGTT
 3' - CCGACTGCAGTCGTT
 3' - CCGACTTCAGTCGTT

5/57

n corresponding nucleotide
 A C T G T T A G C T A A T T G G Ref. Seq.
 C A A \boxed{T} C G A — Probe from first set
 C A A $\boxed{}$ C G A \boxed{T} — Deletion probe
 C A A \boxed{T} A C G \boxed{A} } insertion
 C A A \boxed{T} C C G \boxed{A} } probes
 C A A \boxed{T} G C G \boxed{A} }
 C A A \boxed{T} T C G \boxed{A} }

Fig. 6

8/57

α_1 α_2 α_3 Corresponding nucleotides
 A C T G T T A G C T A A T T G G Ref. Seq.

C \boxed{A} A \boxed{T} C \boxed{G} A Probe from first set
 I_1 I_2 I_3 Interrogation positions

C \boxed{C} A T C G A
 C \boxed{G} A T C G A
 C \boxed{T} A T C G A
 I_1

} Corresponding probes from second, third and fourth probe sets

C A A \boxed{A} C G A
 C A A \boxed{C} C G A
 C A A \boxed{G} C G A
 I_2

} Corresponding probes from fifth, sixth and seventh probe sets

C A A T C \boxed{A} A
 C A A T C \boxed{C} A
 C A A T C \boxed{T} A
 I_3

} Corresponding probes from eighth, ninth and tenth probe sets

Fig. 7

7/57

n_3 n_4 n_1 n_2
 A C T G T T A G C T A A T T G G Ref. Seq.

C A A T C A A T
 C A C T C C A T
 C A G T C G A T
 C A T T C T A T

I_1 I_2 Interrogation positions

T G A C T A T
 T G C C G A T
 T G G C C A T
 T G T C A A T

I_3 I_4 Interrogation positions

Fig. 8

n corresponding nucleotide
 A T T C C C G G G A T C

A G G G C C A T — Probe from first probe set
 A G G C C C A T
 A G G A C C A T
 A G G T C C A T } Corresponding probes from second, third and fourth probe set

$\frac{+}{/}$ $\frac{|}{|}$ hetero mutation

Interrogation
 position

Fig. 9

3157

HV 407A

130 x 140

13/7

15/9

17/9

19/10

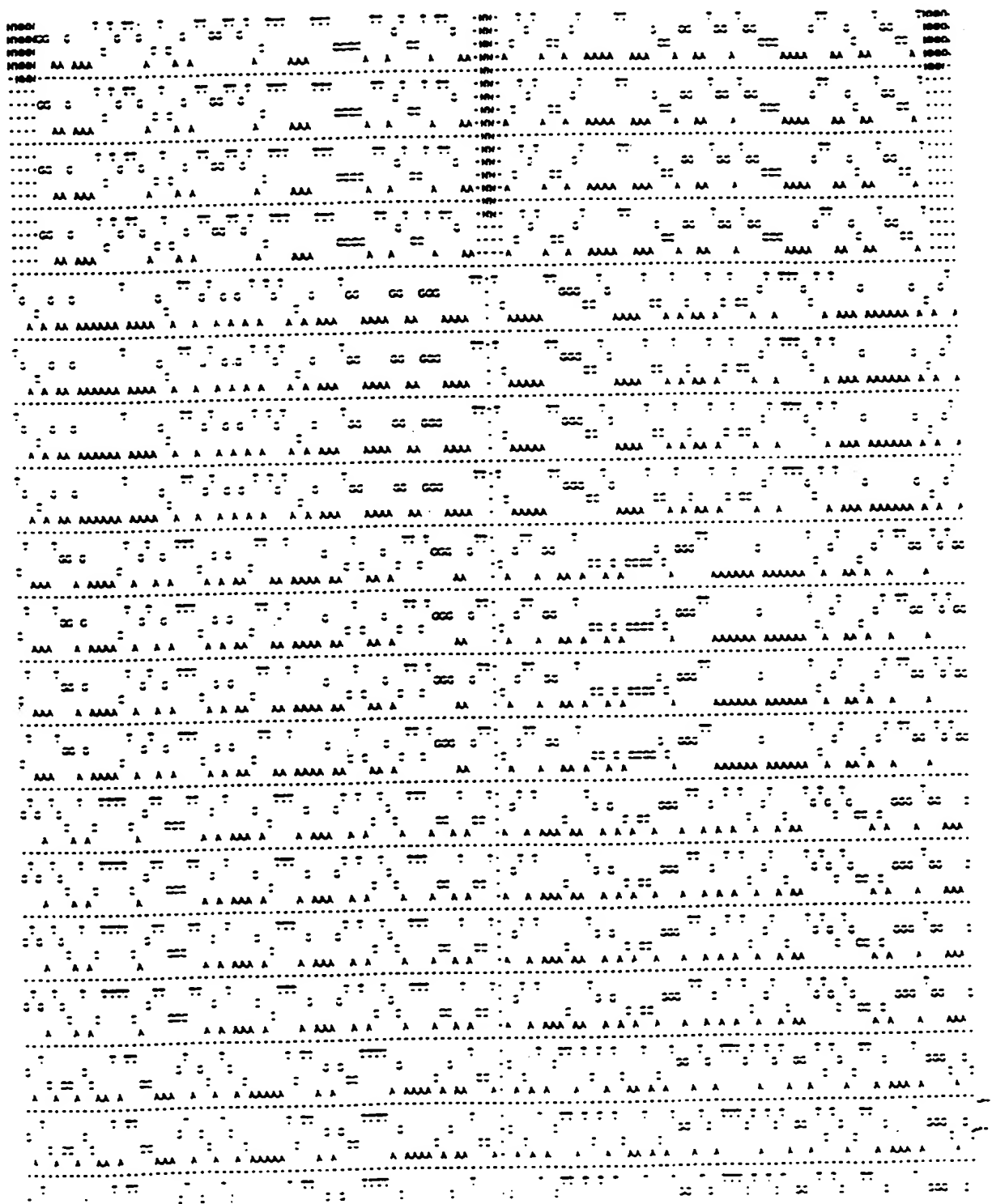


Fig. 10
Page 1 of 2

9/57

Hv7c7A (2)

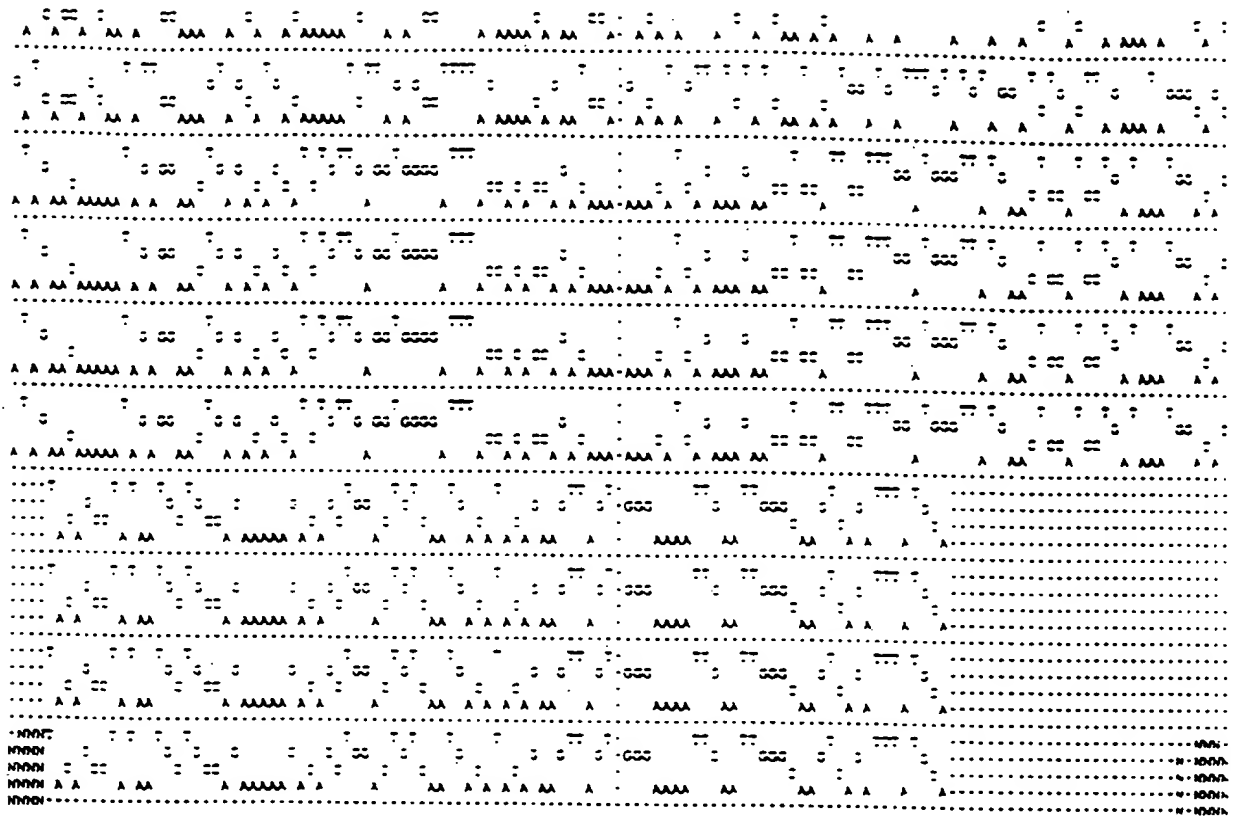
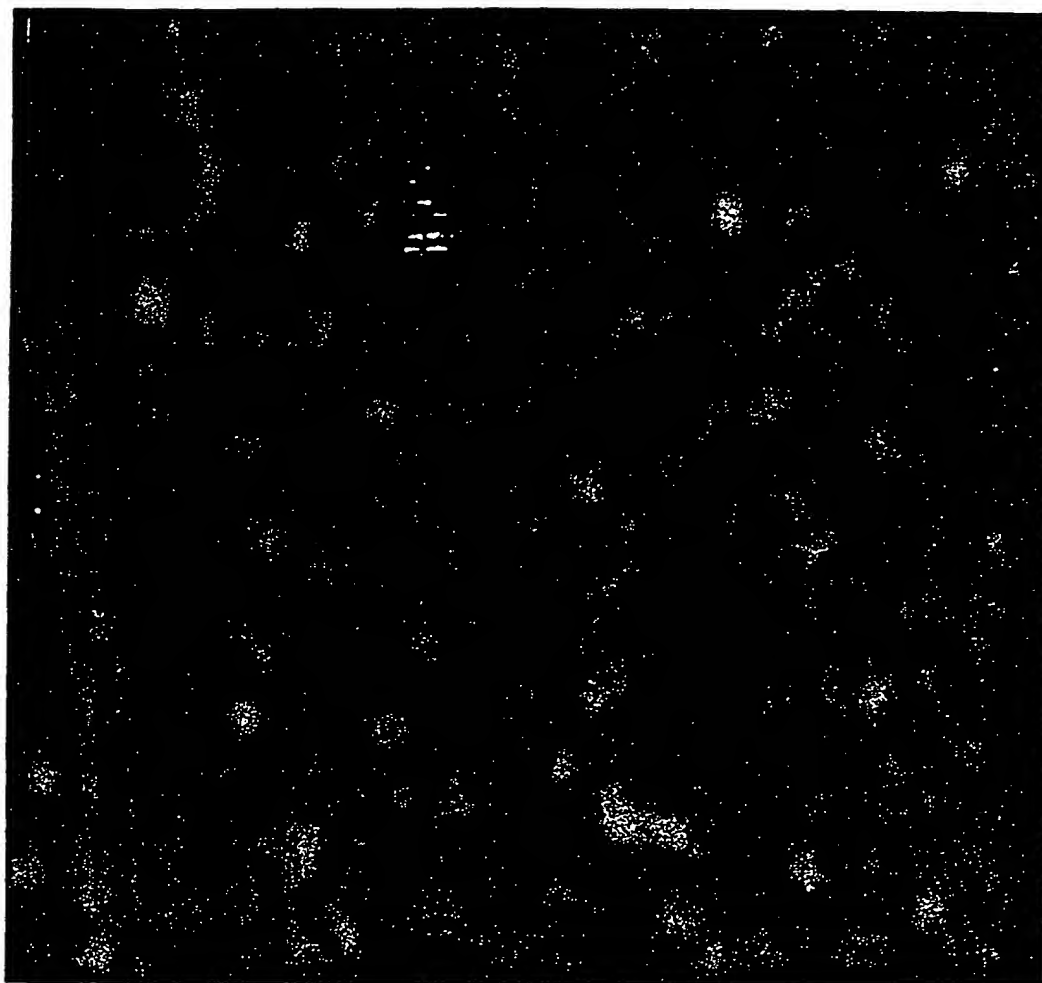


Fig. 10
Page 2 of 2

10.57



13 probe length
15
17
19

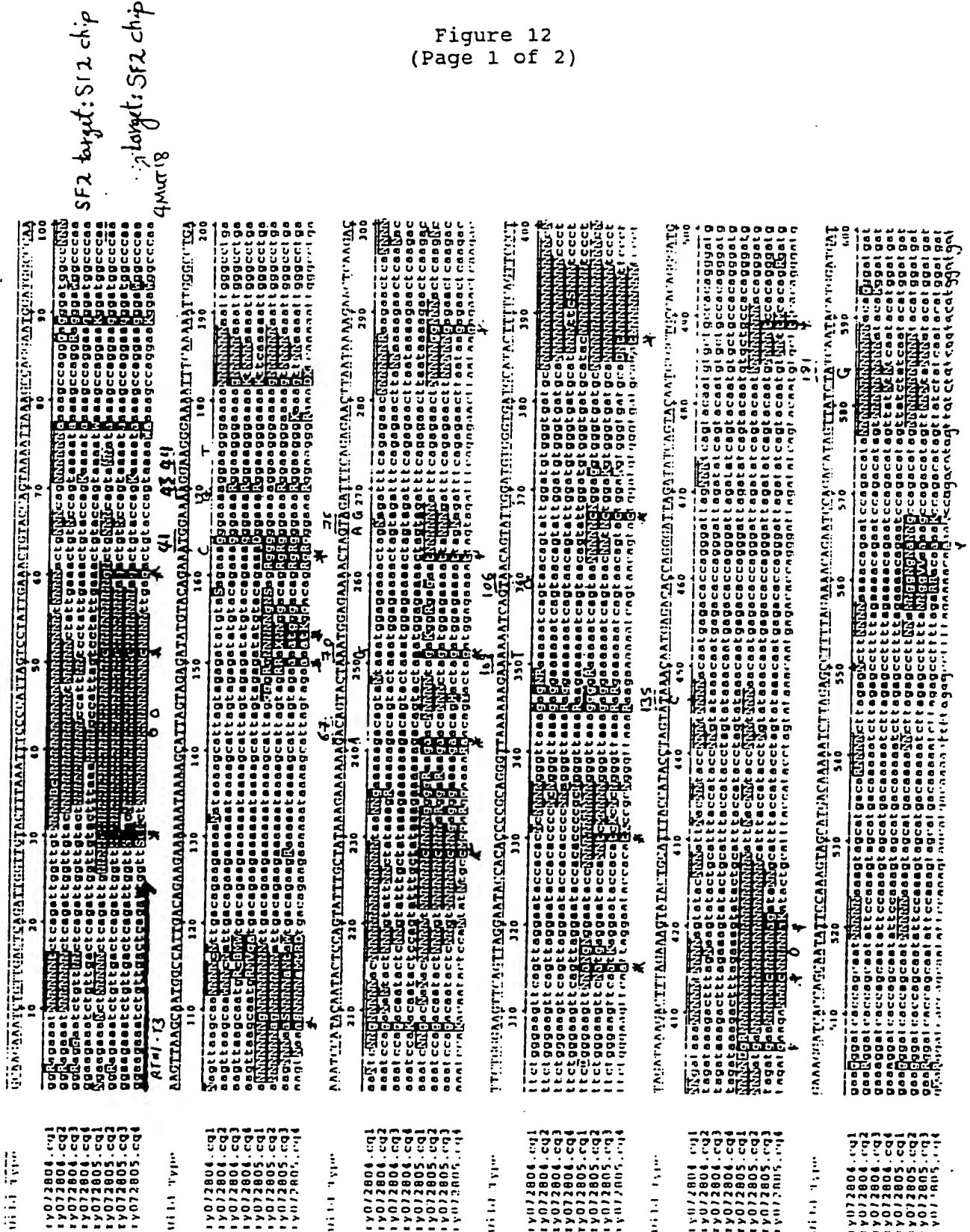
MC07060:

= 407 water chip hybridized with fragmented pPol 19 RNA

Fig. 11

11/87

Figure 12
(Page 1 of 2)



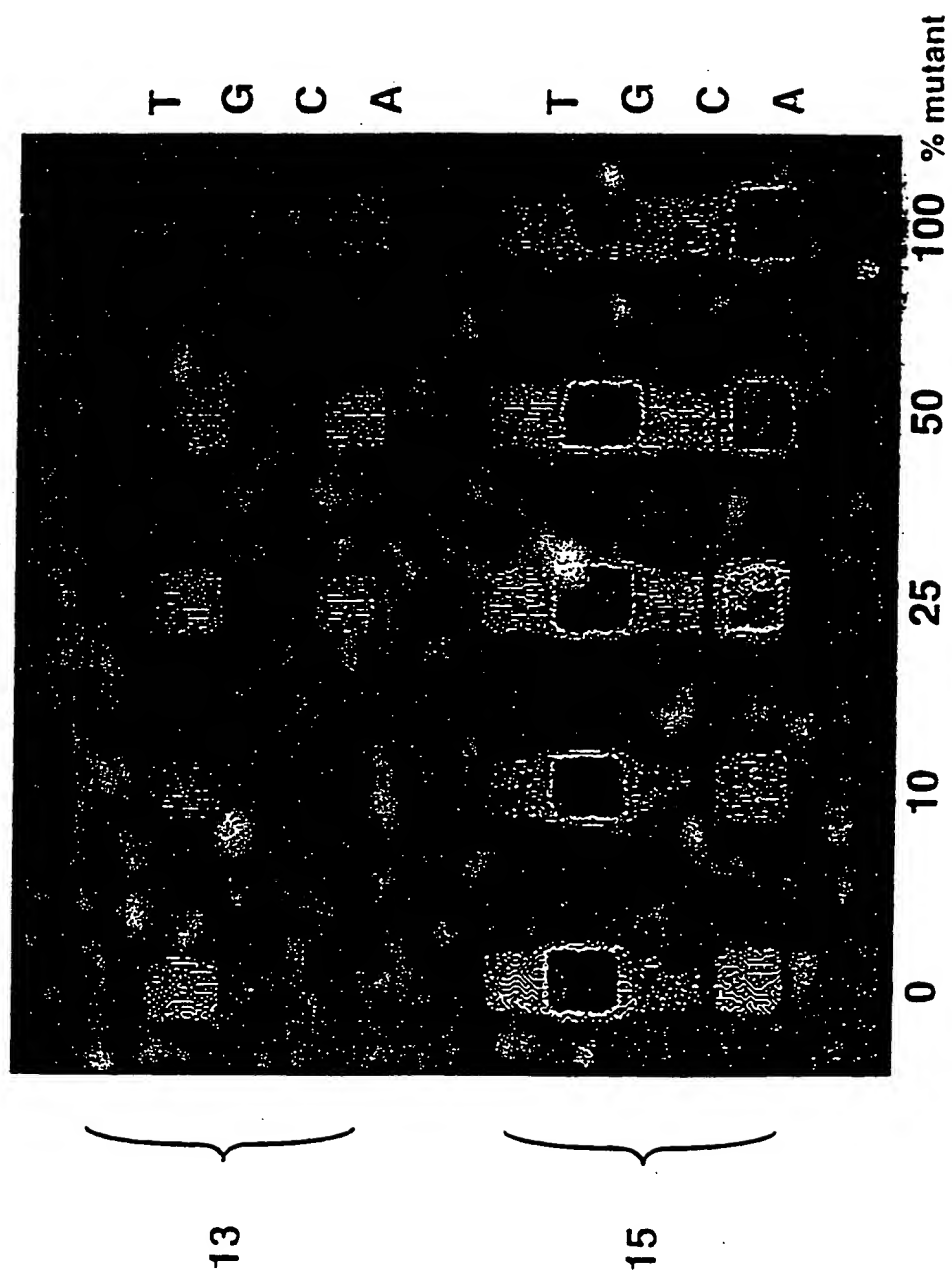
13/57

5'Fluorescein-AAAGAAAAAAGACAGTACTAAATGGAGAAAAT wildtype
PROBE 3' tttttt•tgtcat 13mers
PROBE 3' cttttttt•tgtcatg 15mers
PROBE 3' tcttttttt•tgtcatga 17mers
PROBE 3' ttcttttttt•tgtcatgat 19mers
5'Fluorescein-AAAGAAAAAACAAGTACTAAATGGAGAAAAT mutant

Fig. 13

14/57

Fig. 14



10 pre and post-addict treated Patients

[illegible]

ms093002.002	176	177	178	179	180	181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200	201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220	221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280	281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300	301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320	321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340	341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360	361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380	381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400	401	402	403	404	405	406	407	408	409	410	411	412	413	414	415	416	417	418	419	420	421	422	423	424	425	426	427	428	429	430	431	432	433	434	435	436	437	438	439	440	441	442	443	444	445	446	447	448	449	450	451	452	453	454	455	456	457	458	459	460	461	462	463	464	465	466	467	468	469	470	471	472	473	474	475	476	477	478	479	480	481	482	483	484	485	486	487	488	489	490	491	492	493	494	495	496	497	498	499	500	501	502	503	504	505	506	507	508	509	510	511	512	513	514	515	516	517	518	519	520	521	522	523	524	525	526	527	528	529	530	531	532	533	534	535	536	537	538	539	540	541	542	543	544	545	546	547	548	549	550	551	552	553	554	555	556	557	558	559	560	561	562	563	564	565	566	567	568	569	570	571	572	573	574	575	576	577	578	579	580	581	582	583	584	585	586	587	588	589	590	591	592	593	594	595	596	597	598	599	600	601	602	603	604	605	606	607	608	609	610	611	612	613	614	615	616	617	618	619	620	621	622	623	624	625	626	627	
--------------	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	--

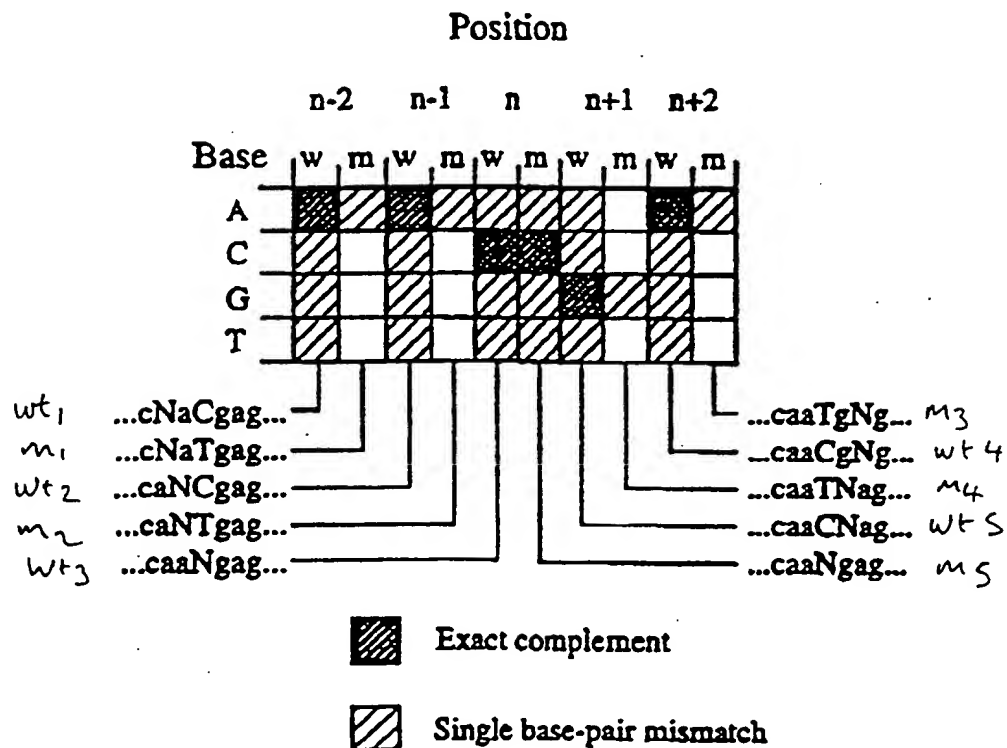
↑
↑↑↑
↑
nucleotido
207

Fig. 15

16/57

Array Design for the R553X Point Mutation

Wild-Type Pattern



Wild-Type Sequence: 5'-AGGTCAA**C**GAGCAA-3'

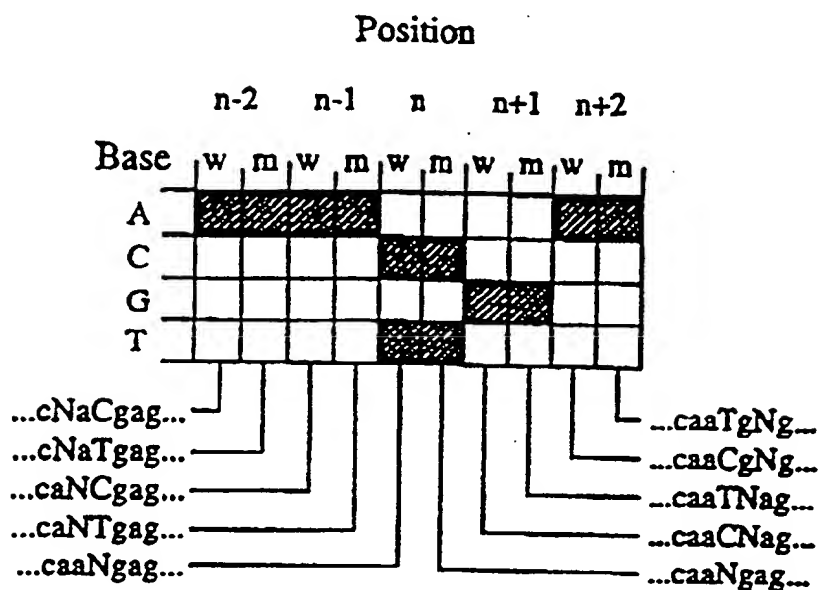
Mutant Sequence: 5'-AGGTCAA**T**GAGCAA-3'

Fig. 16

17/57

Array Design for the R553X Point Mutation

Heterozygote Pattern

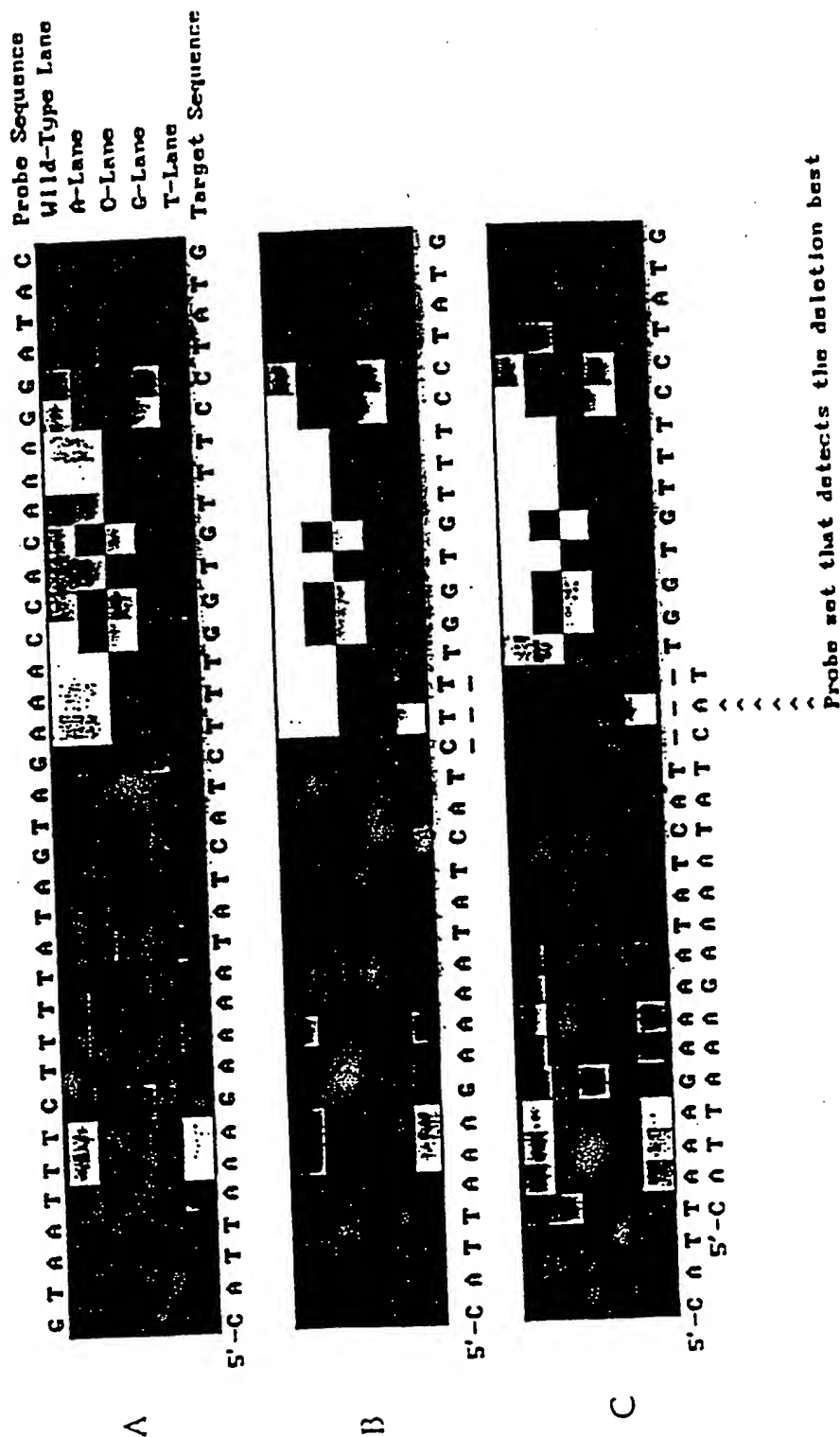


Wild-Type Sequence: 5'-AGGTCAA**C**GAGCAA-3'

Mutant Sequence: 5'-AGGTCAA**T**GAGCAA-3'

Fig. 17

13 / 57



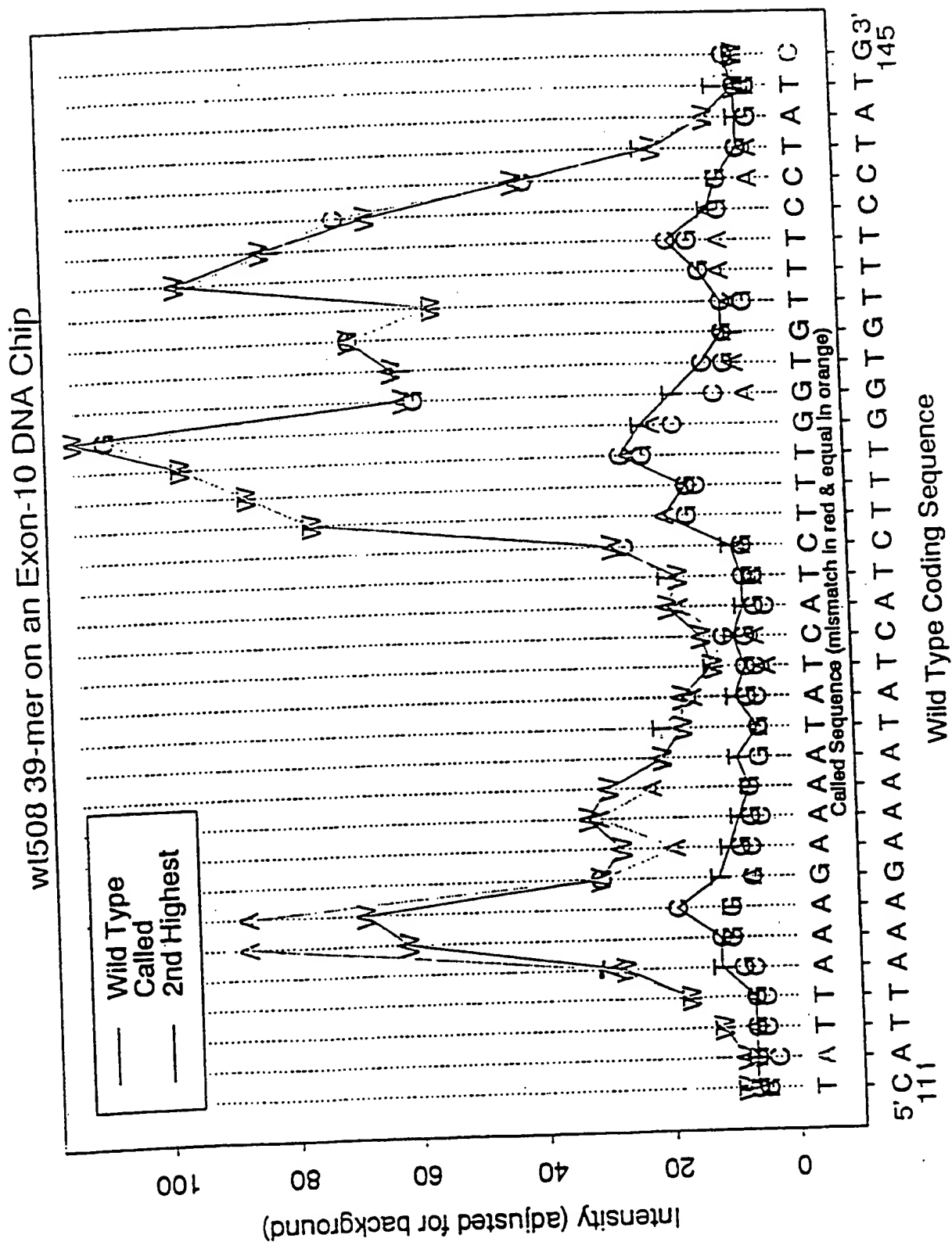


Fig. 19
Page 1 of 3

20/57

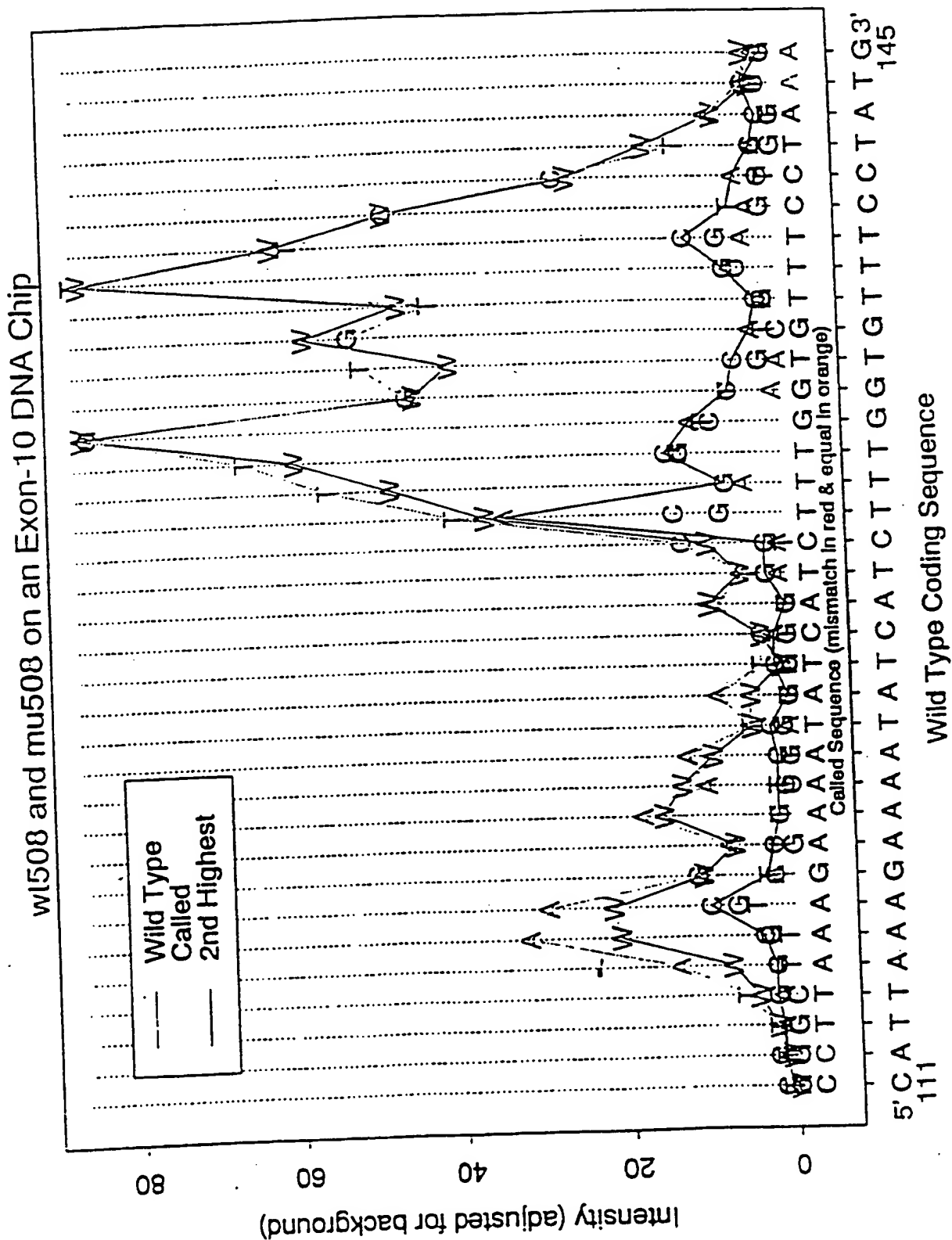


Fig. 19
Page 2 of 3

21,57

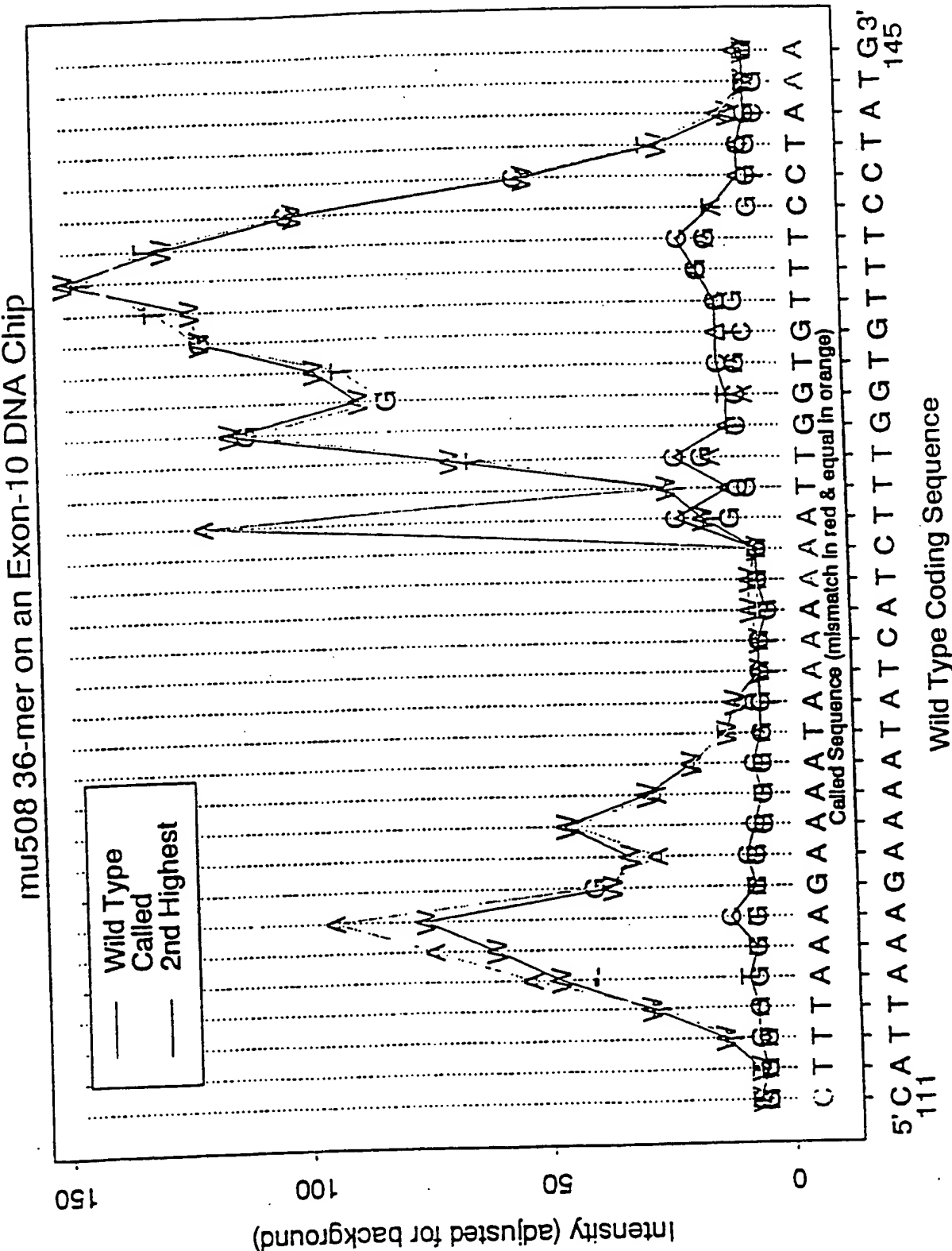


Fig. 19
Page 3 of 3

22/57

Probe Sequence
Wild-Type Lane
A-Lane
C-Lane
G-Lane
T-Lane
Target Sequence

GGAGTCTCCCATTTAATT
5'-CCTTCAGAGGGTAAATTA

A

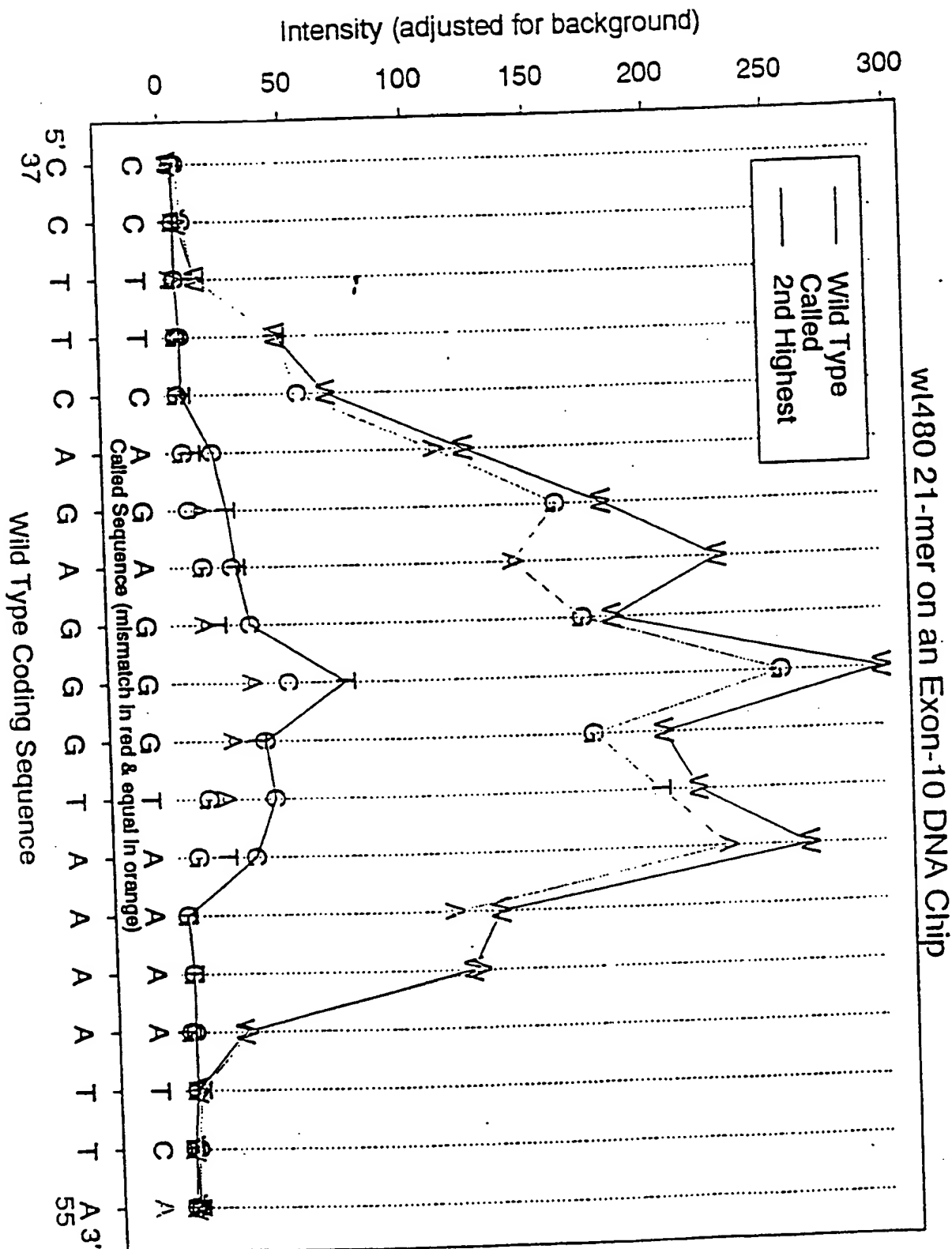
5'-CCTTCAGAGGGTAAATTA

B

5'-CCTTCAGAGTGTAAATTA

C

Fig. 20



23/07

24/57

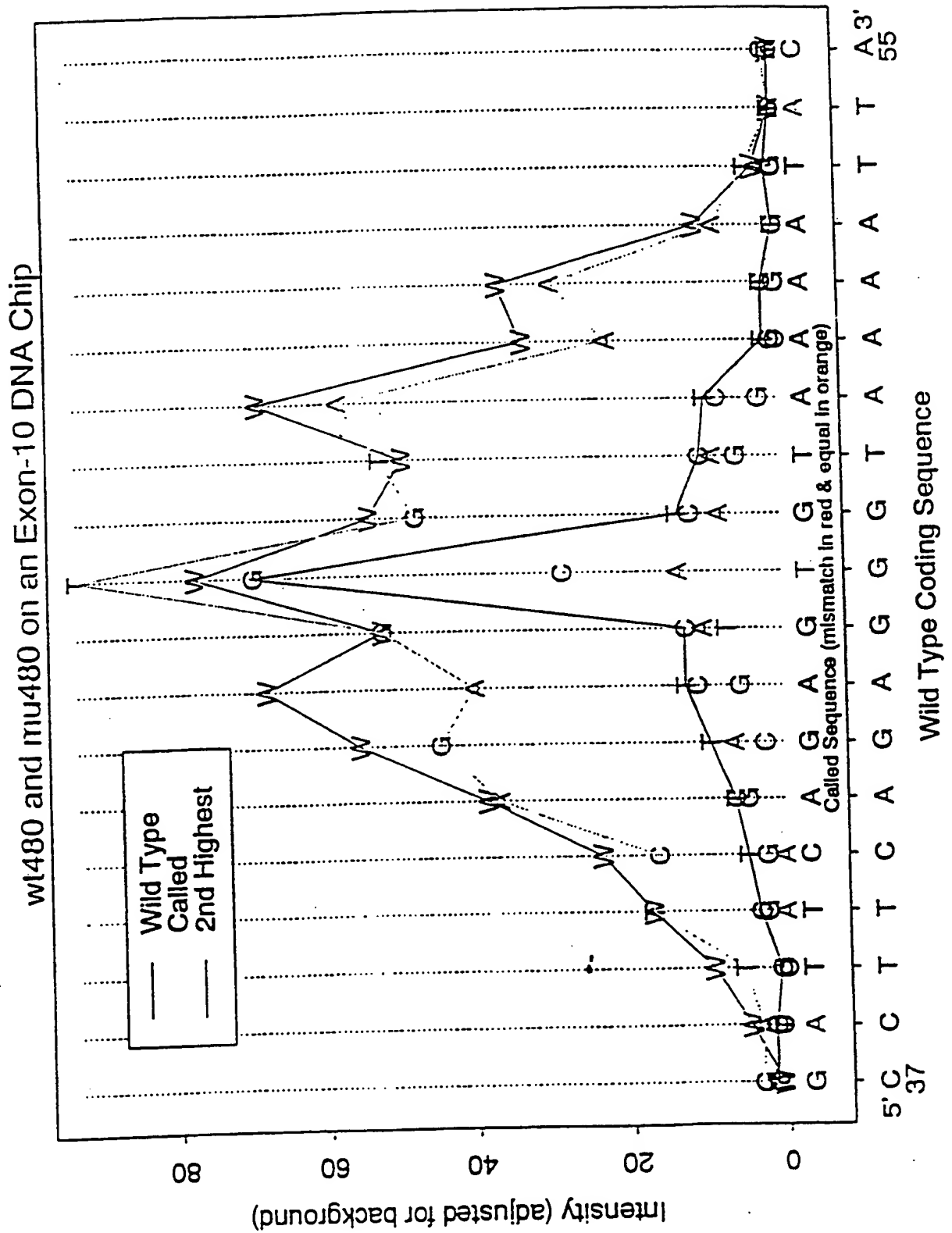


Fig. 21
Page 2 of 3

25 / 57

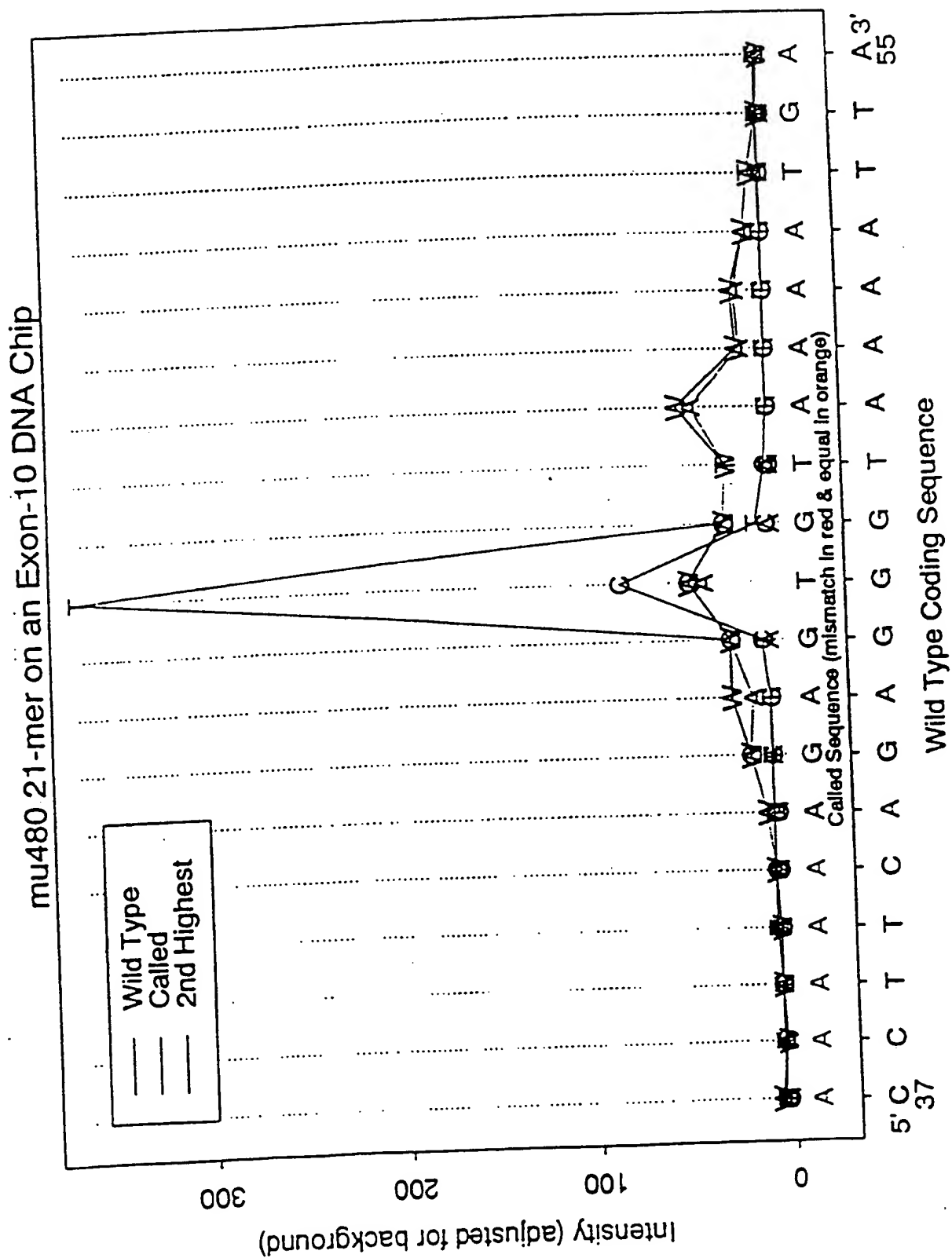


Fig. 21
Page 3 of 3

26/57

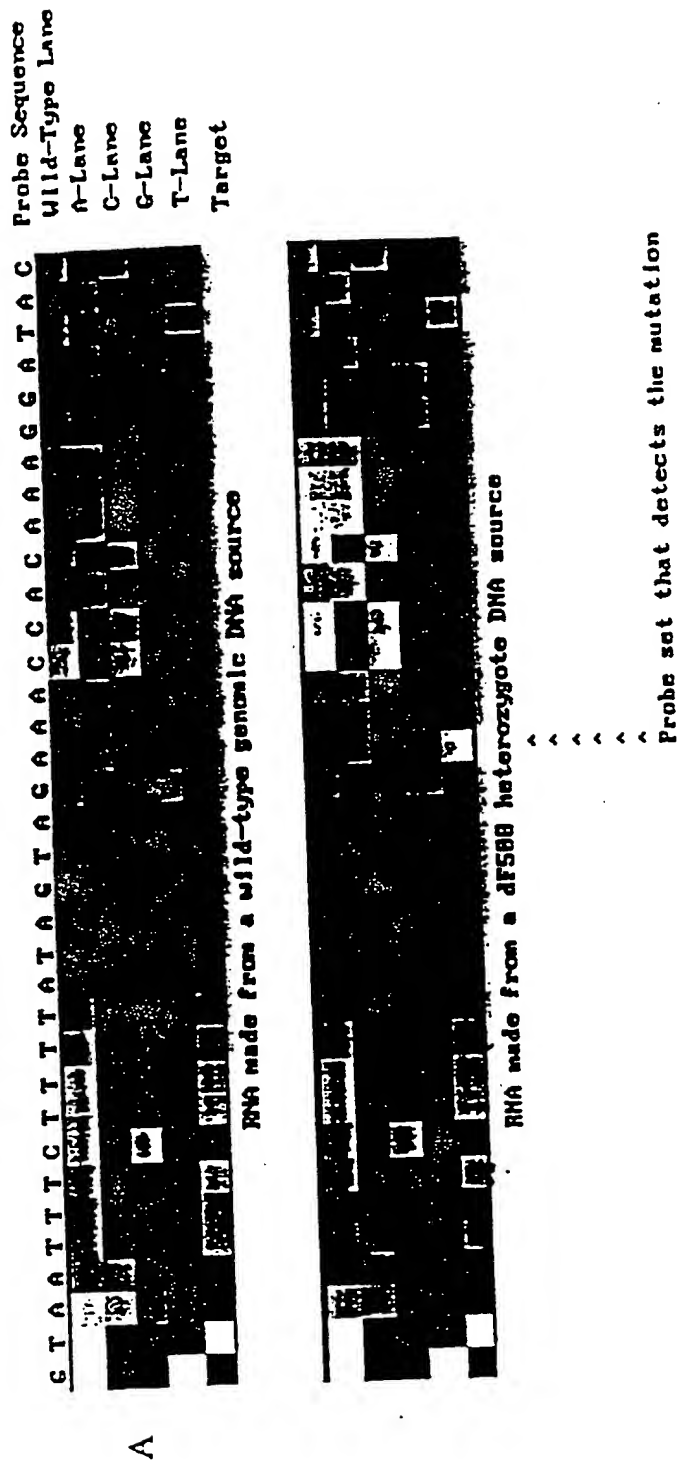


Fig. 22

27/57

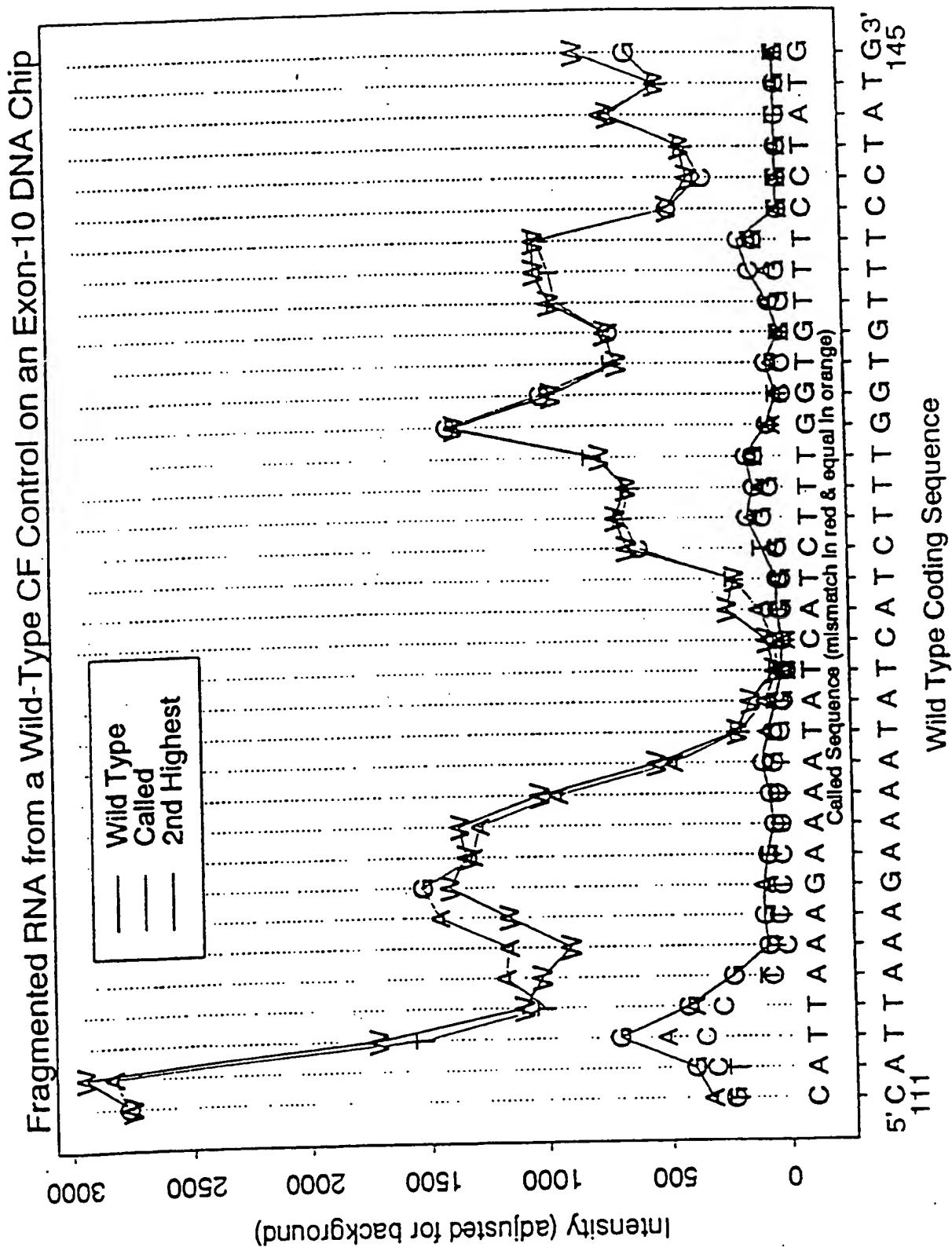


Fig. 23
Page 1 of 2

28/57

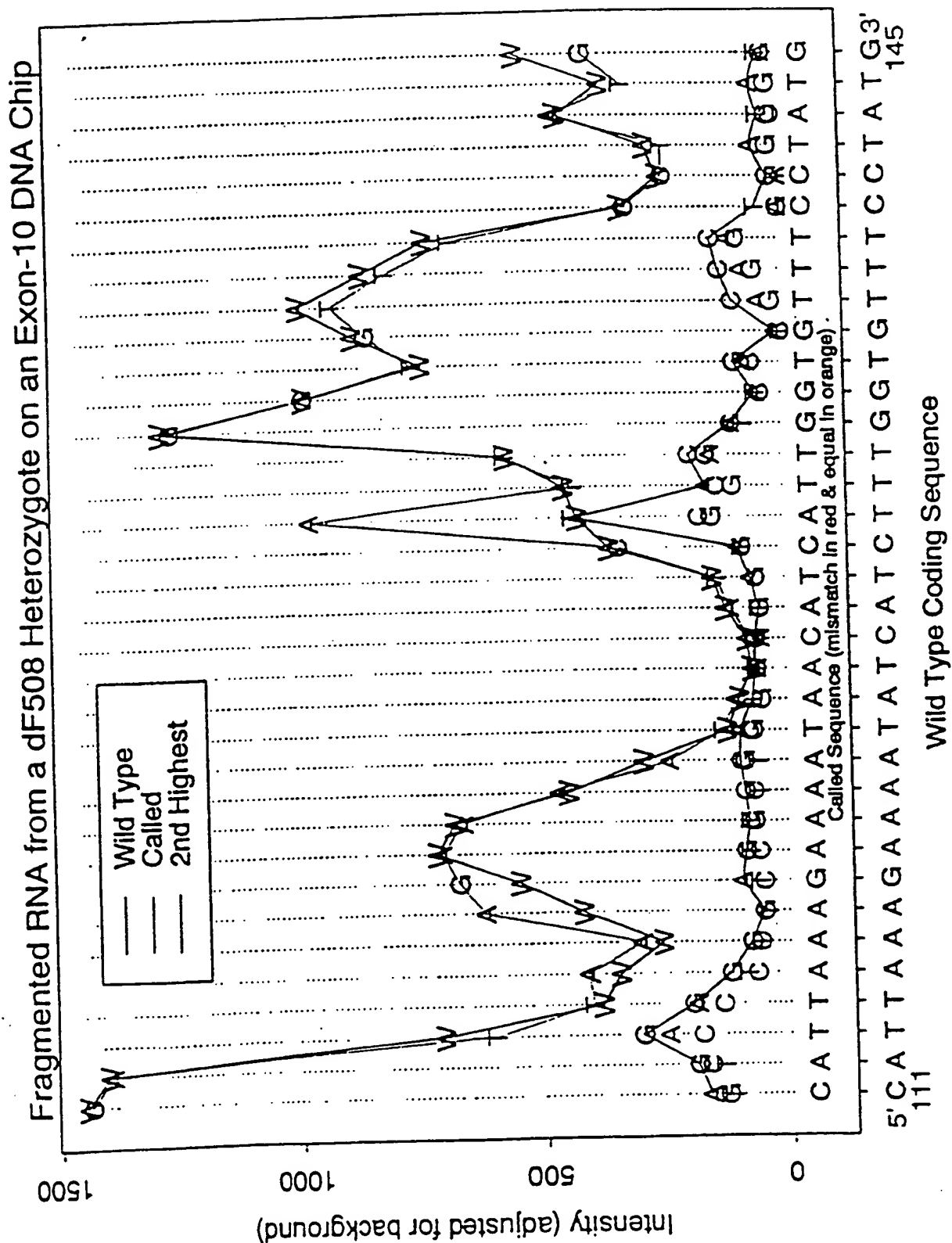


Fig. 23

20/57.

A

B

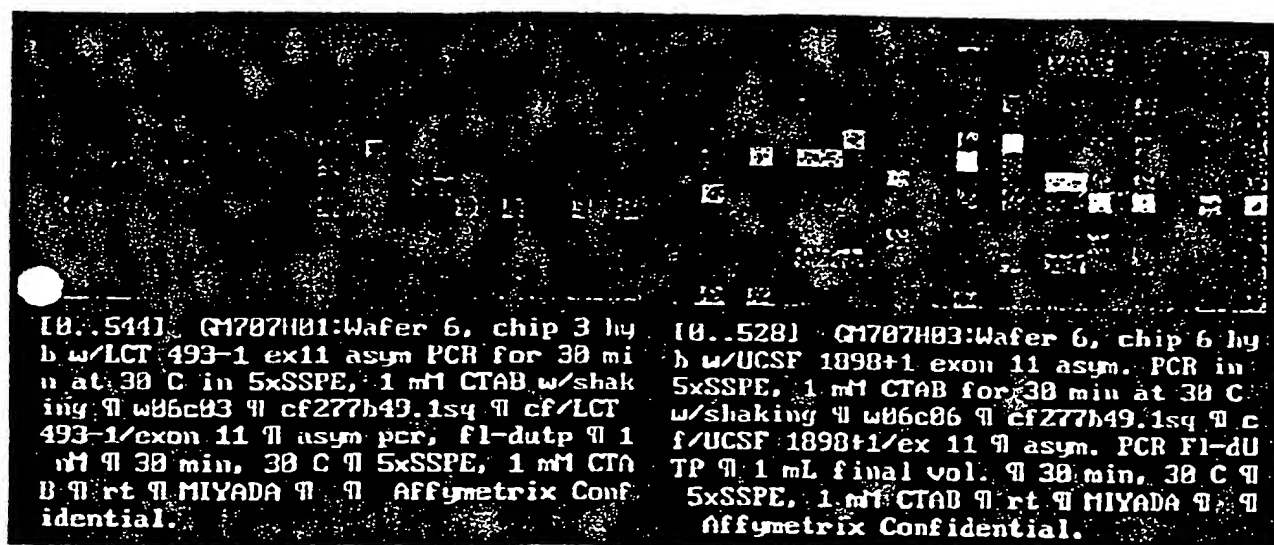


Fig. 24

80.17

A

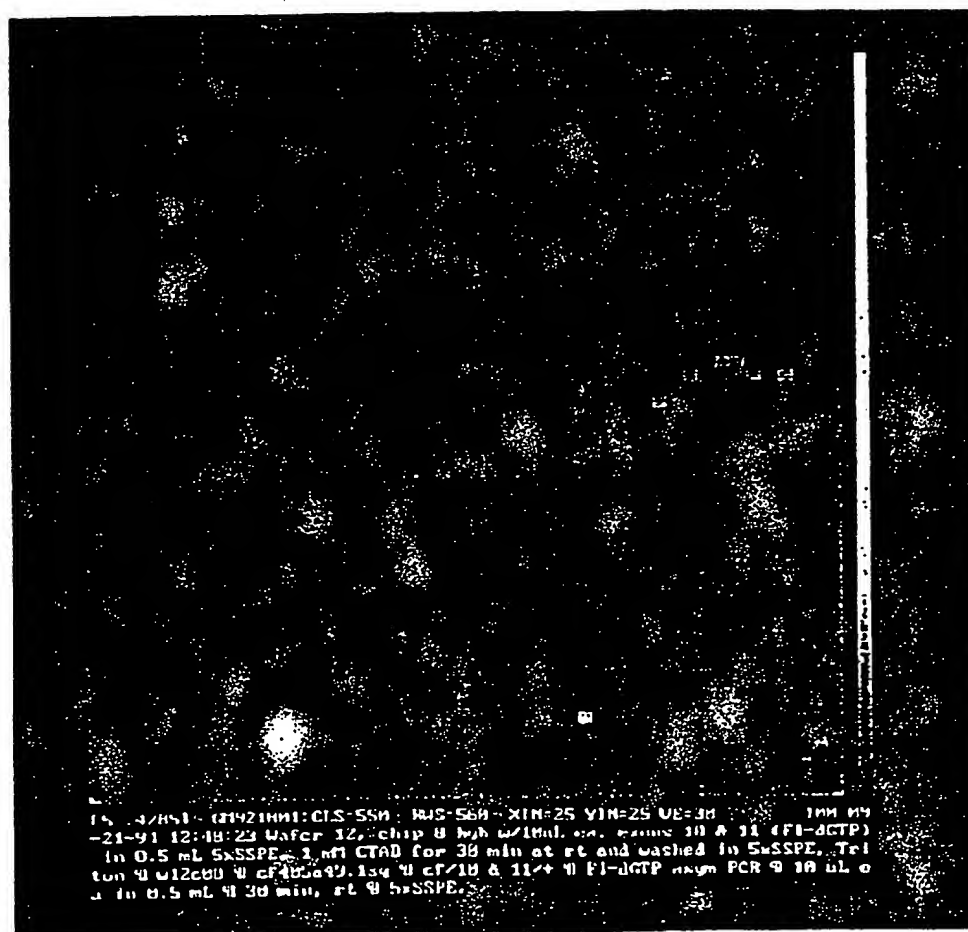


Fig. 25
Page 1 of 2

31/57

B

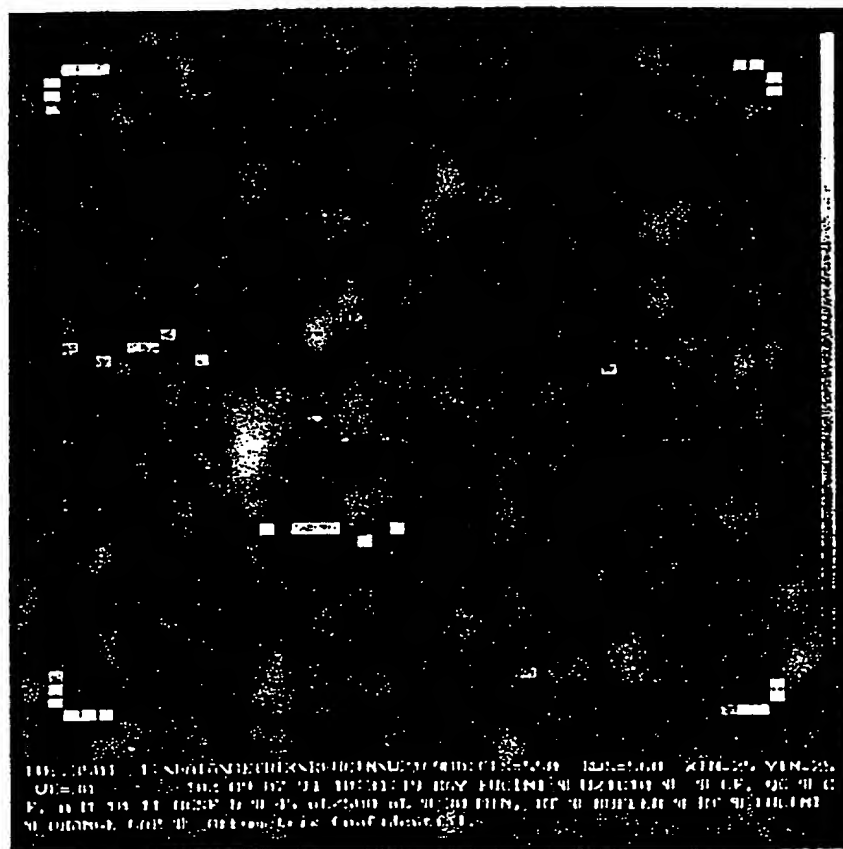


Fig. 25
Page 2 of 2

32/57

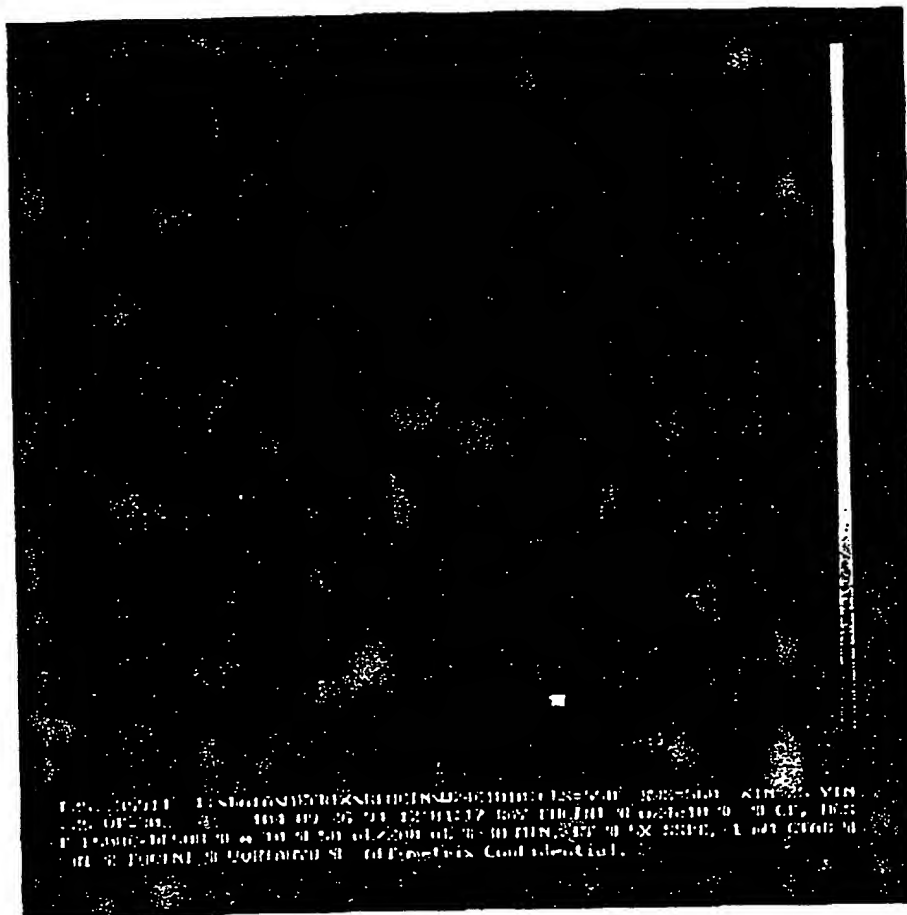


Fig. 26

33/57

P53 EXON 6 CODON 192 REGION: 12MER PROBES

0.	G	A	T	G	C	T	G	A	G	G	G	T
1.				A			C	T	C	C	G	G
2.				G	A		C	T	C	C	G	G
3.				C	G	A	C	T	C	C	C	C
4.				A	C	G	A	T	C	C	C	C
5.				T	A	C	A	T	C	C	C	C
6.				C	T	A	C	T	C	C	C	C
7.				T	T	A	C	T	C	C	C	C
8.				T	T	A	C	T	C	C	C	C
9.				A	T	A	C	T	C	C	C	C
10.				T	A	A	C	T	C	C	C	C
11.				C	T	A	C	T	C	C	C	C
12.				C	C	A	C	T	C	C	C	C

Fig. 27

35/57.

Detection of 12-mer One-Base Substitution P53 Targets

Fig. 31

4:1 Mixture of WT and
"A" Substitution 12-mer
Targets

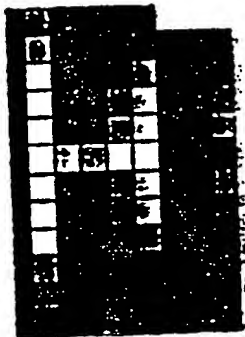
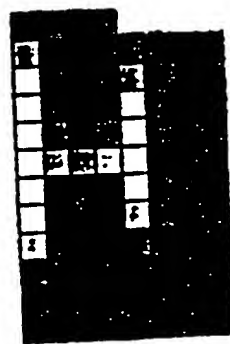
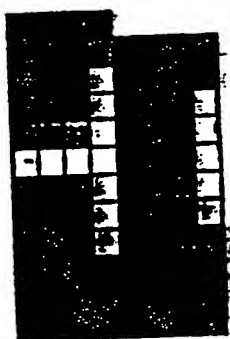


Fig. 29

WT ("G" Substitution)
Target 12-mer



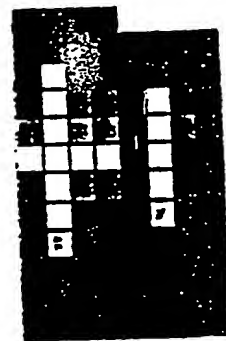
"A" Substitution 12-mer Target



"C" Substitution Target 12-mer



"T" Substitution Target 12-mer



Figs. 29 and 31

36/57

P53 EXON 6 CODON 192 REGION

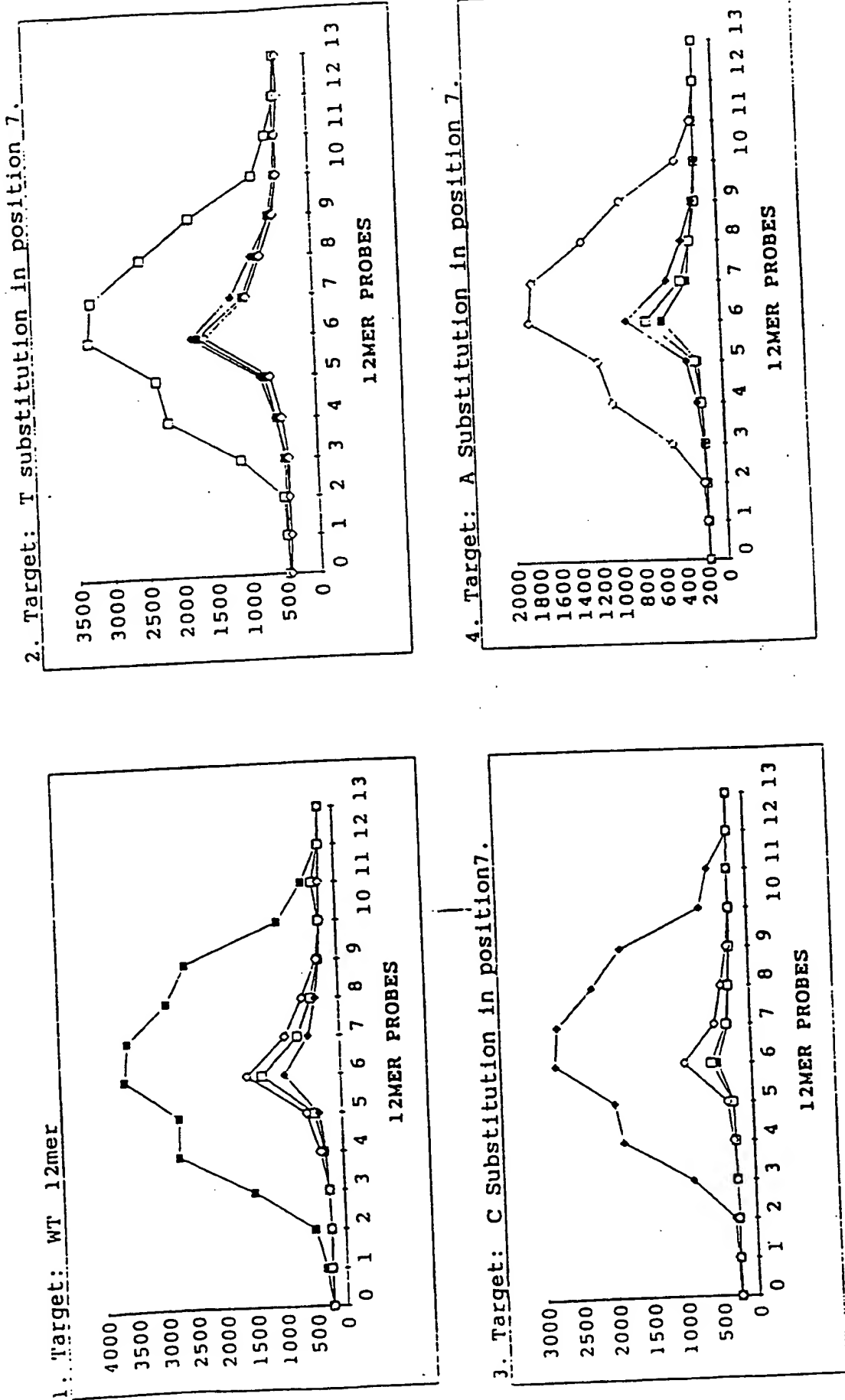


Fig. 30

37/57

P53 EXON 6 CODON 192 REGION

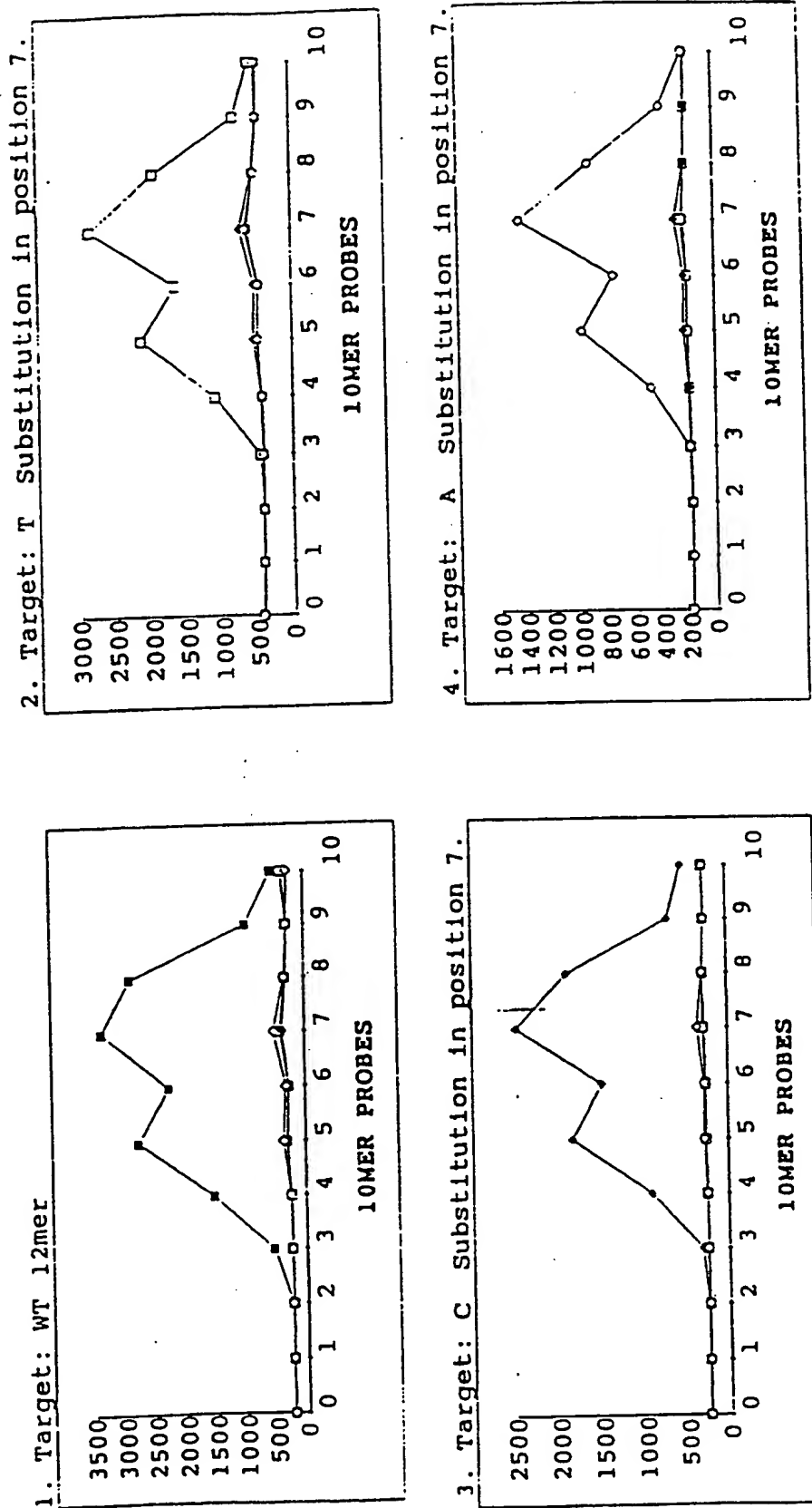


Fig. 32

88/57

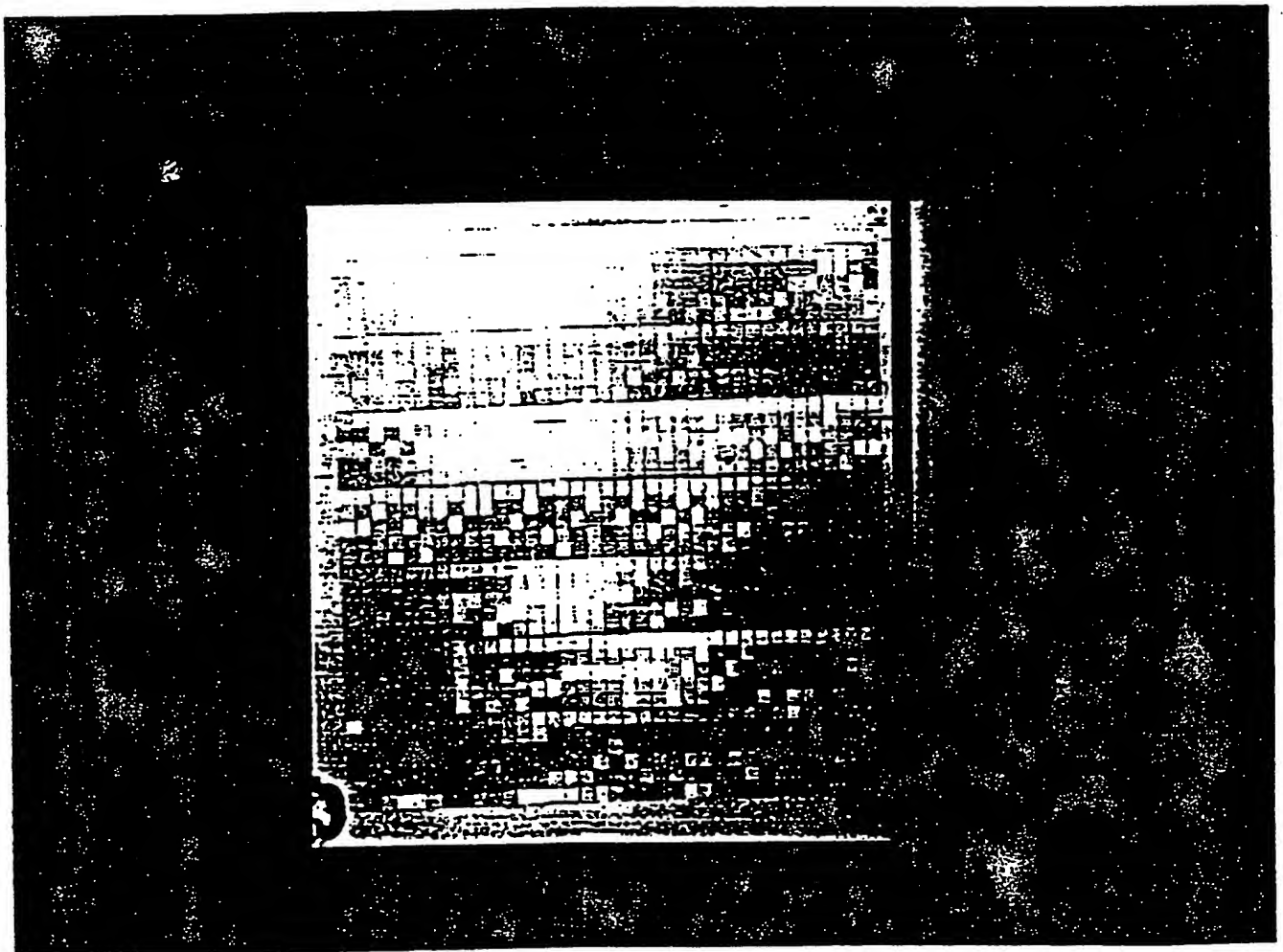


Fig. 33

40/57.

THE HUMAN MITOCHONDRIAL GENOME

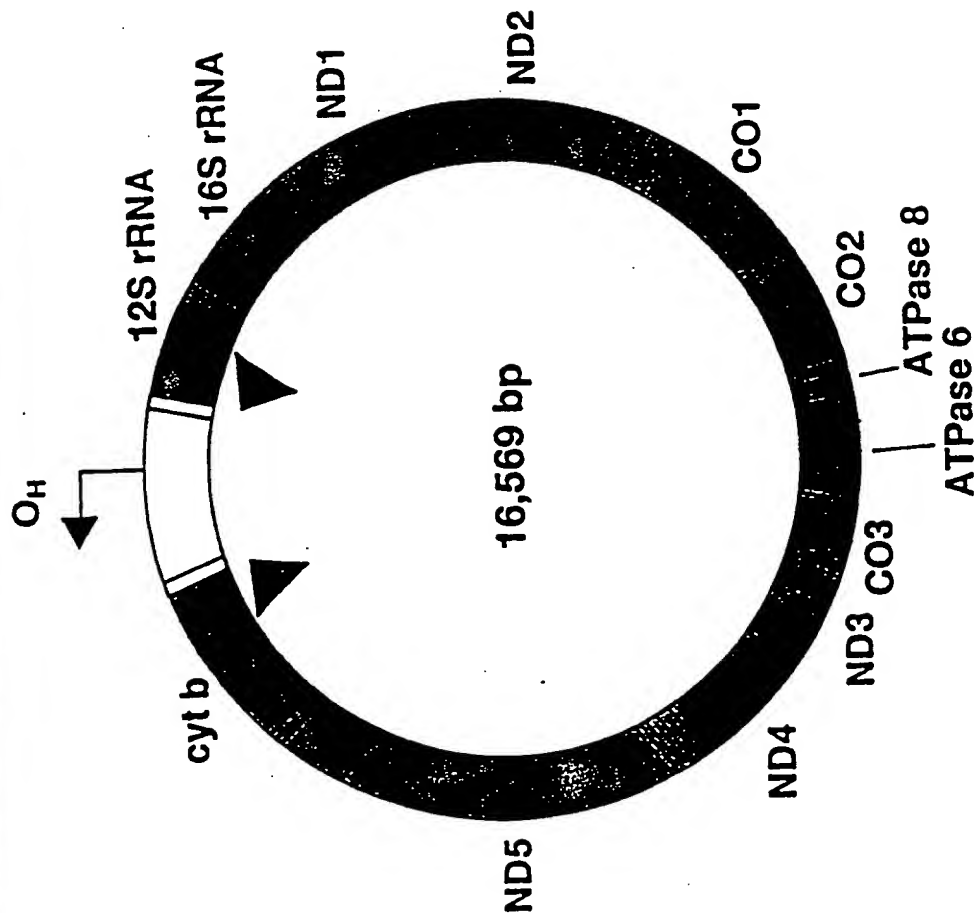
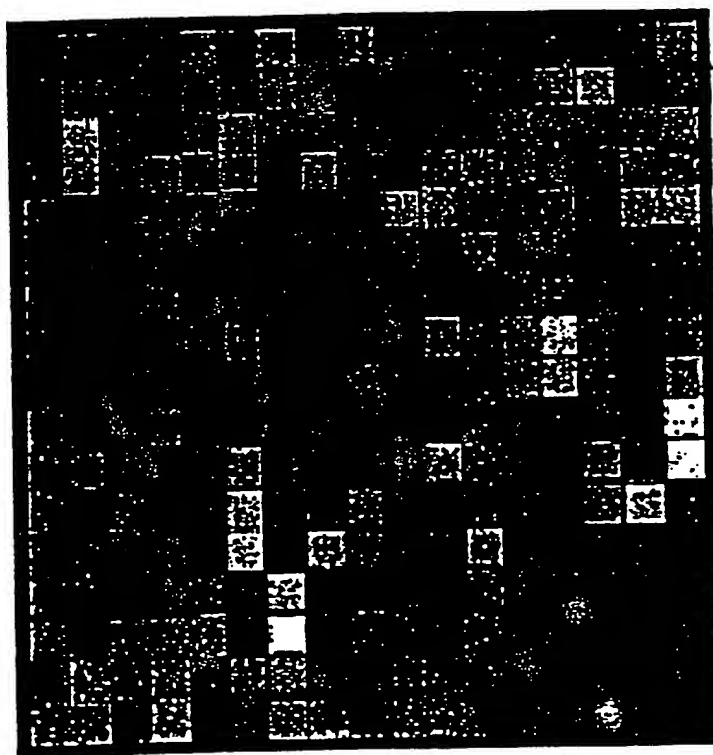


Fig. 35

41/57

mt4

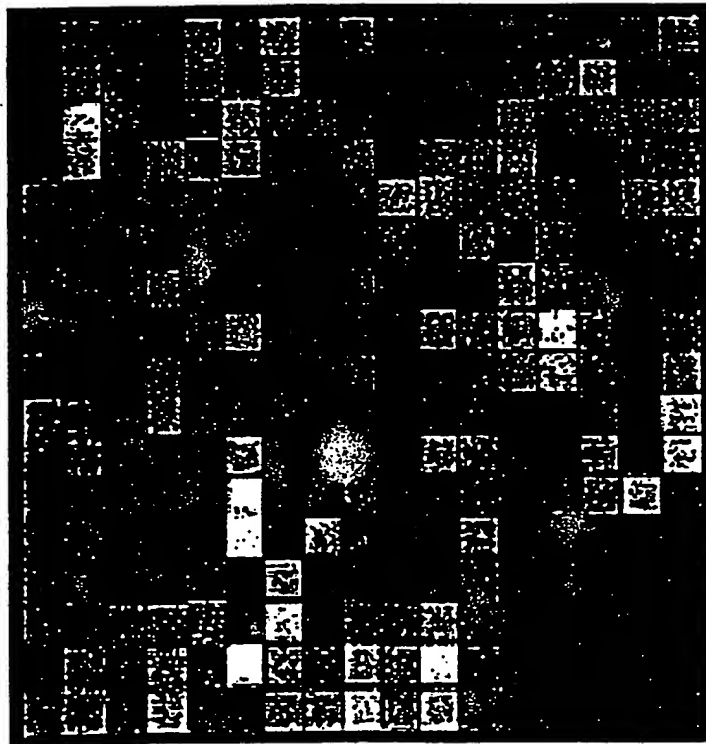


HYBRIDIZATION

Fig. 36

42/57

mt5



HYBRIDIZATION

Fig. 37

43/57

PREDICTED DIFFERENCE IMAGE

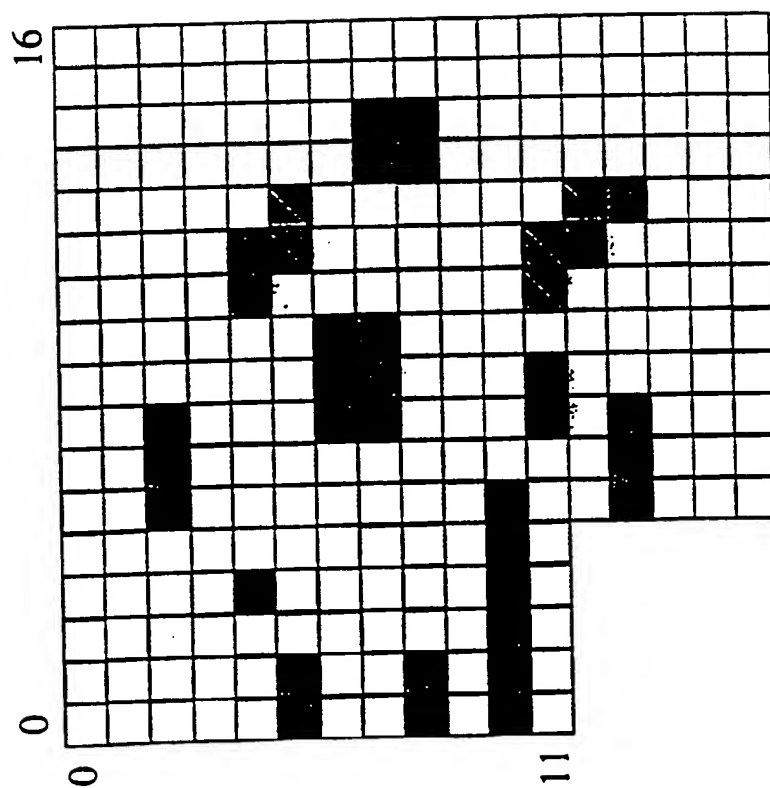


Fig. 38

44/57

DIFFERENCE IMAGE

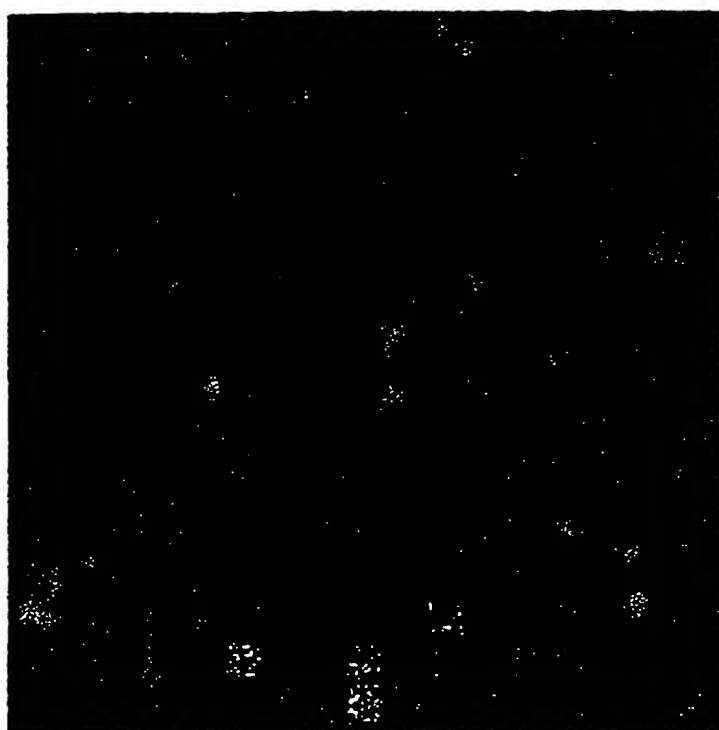


Fig. 39

NORMALIZED INTENSITIES

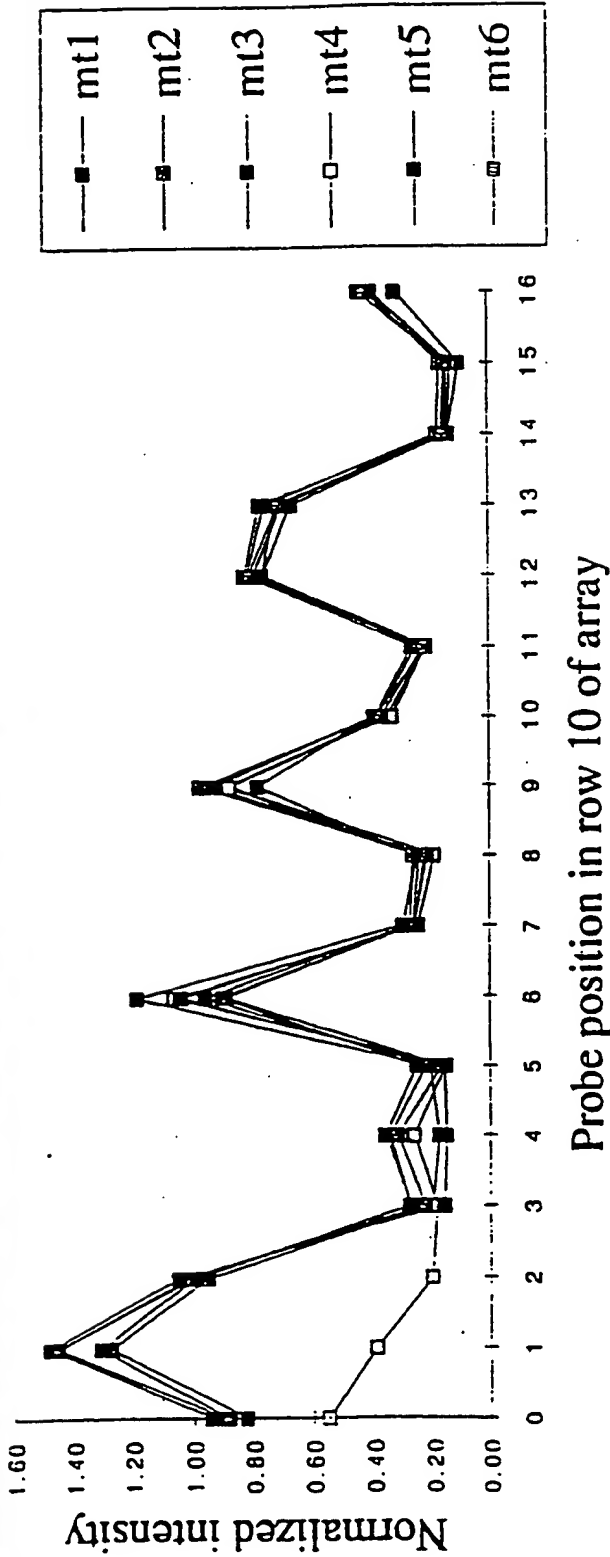


Fig. 40
Sheet 1 of 2

probe position	0	1	2	3	4	5
probe length	13	13	12	12	12	12
sample (mt1 -> 6)	4	4	4	2, 5	2, 5	2, 5
mismatch position from 3' of probe	12	5	3	12	7	2
base change	t -> a	t -> a	t -> a	t -> c	t -> c	t -> c

46/57

NORMALIZED INTENSITIES

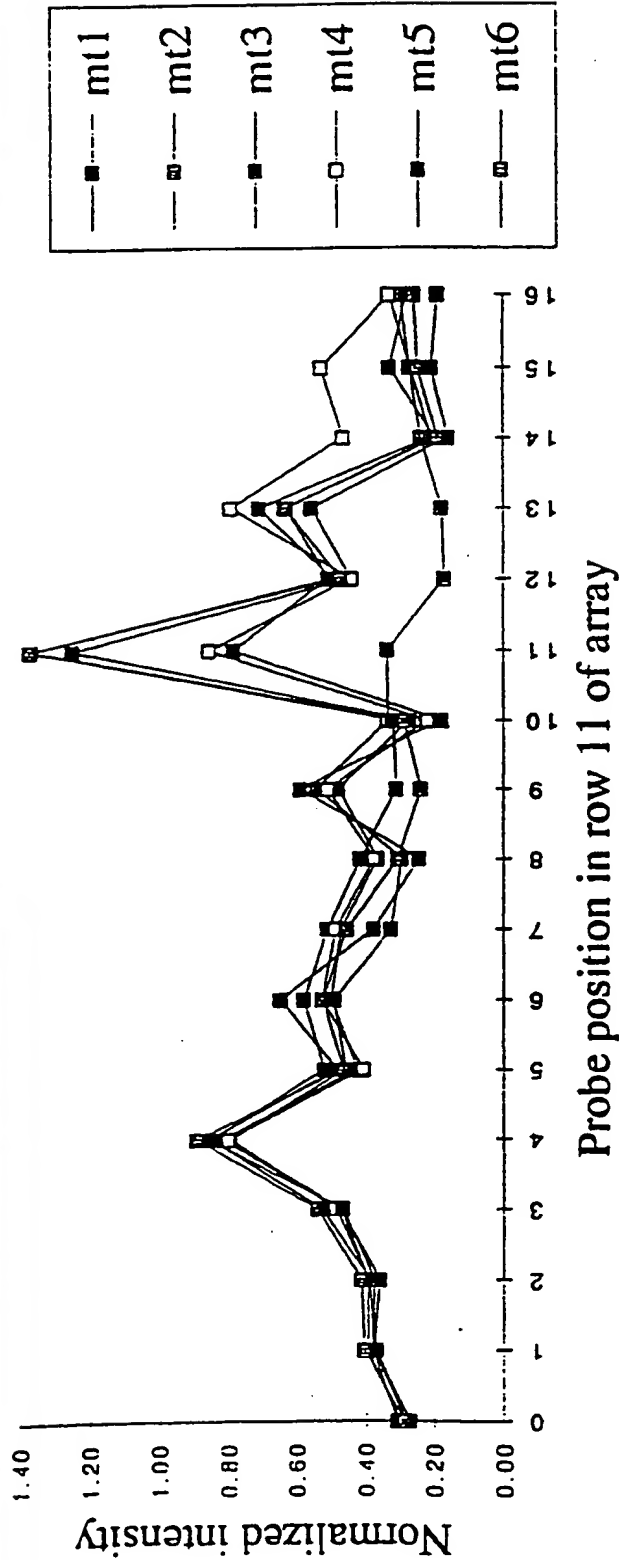


Fig. 40
Sheet 2 of 2

probe position	6	7	8	9	10	11	12	13
probe length	13	12	12	13	14	13	12	12
sample (mt1 -> 6)	2	2, 5	2, 5, 6	3, 6	3, 4, 5	2, 4, 5	2	2
mismatch position from 3' of probe	13	9, 10	3, 4 11	11, 5	4, 11, double	11, 3, double	6	3
base change	c->t	c->t	c->t t->c	t->c	t->c double	g->a t->c double	g->a	g->a

47/57

DISCRIMINATION

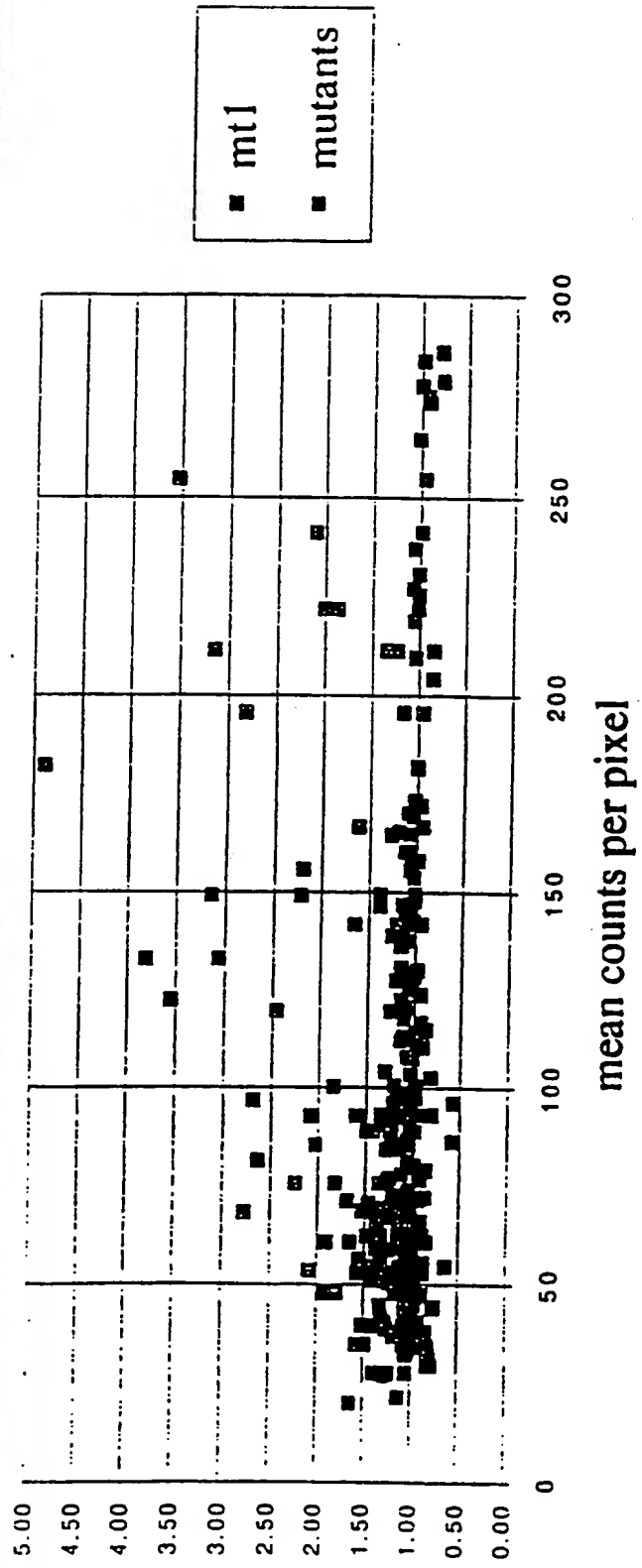


Fig. 41

SEQUENCE & POSITION OF MUTATION

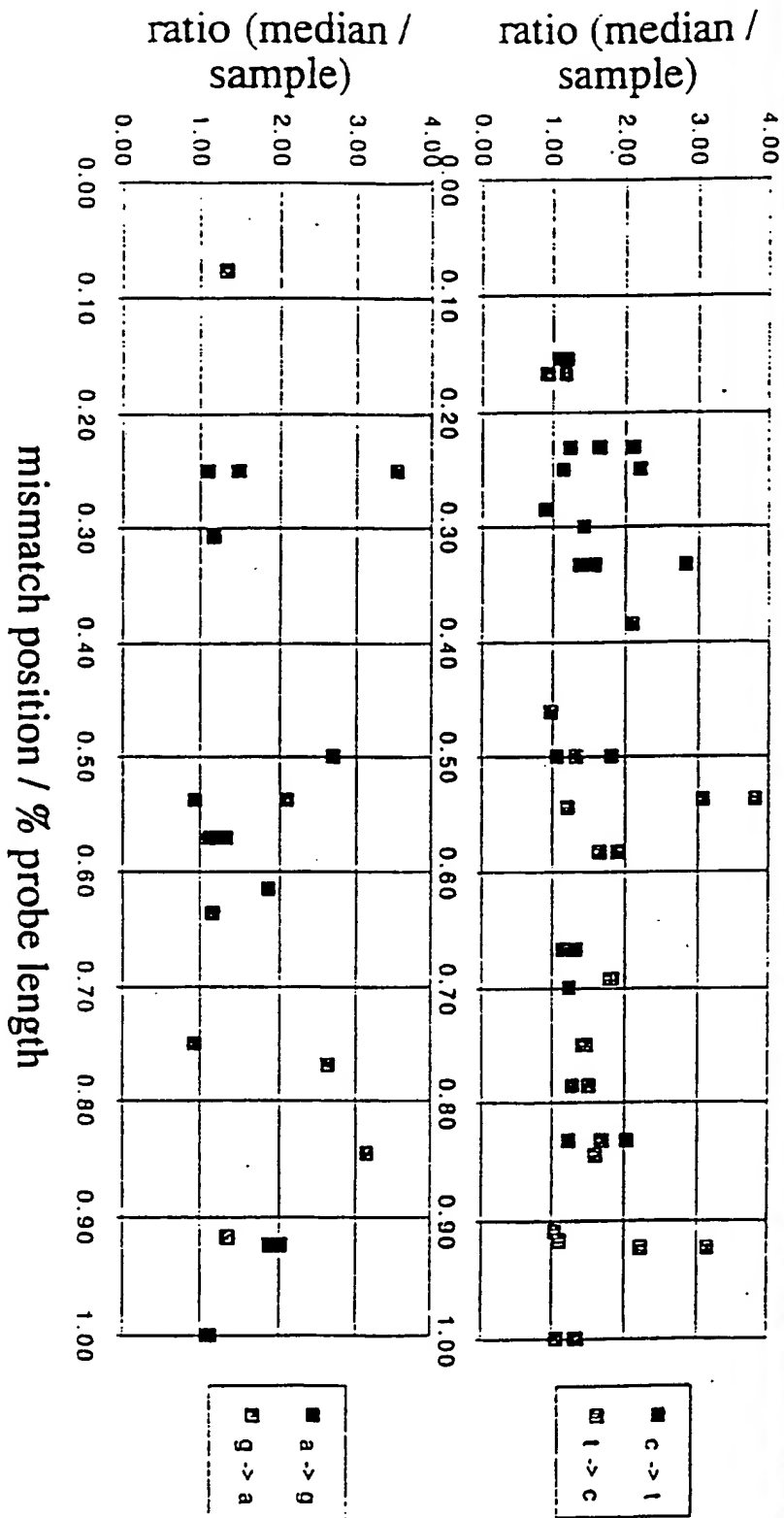


Fig. 42

49/57

SEQUENCE

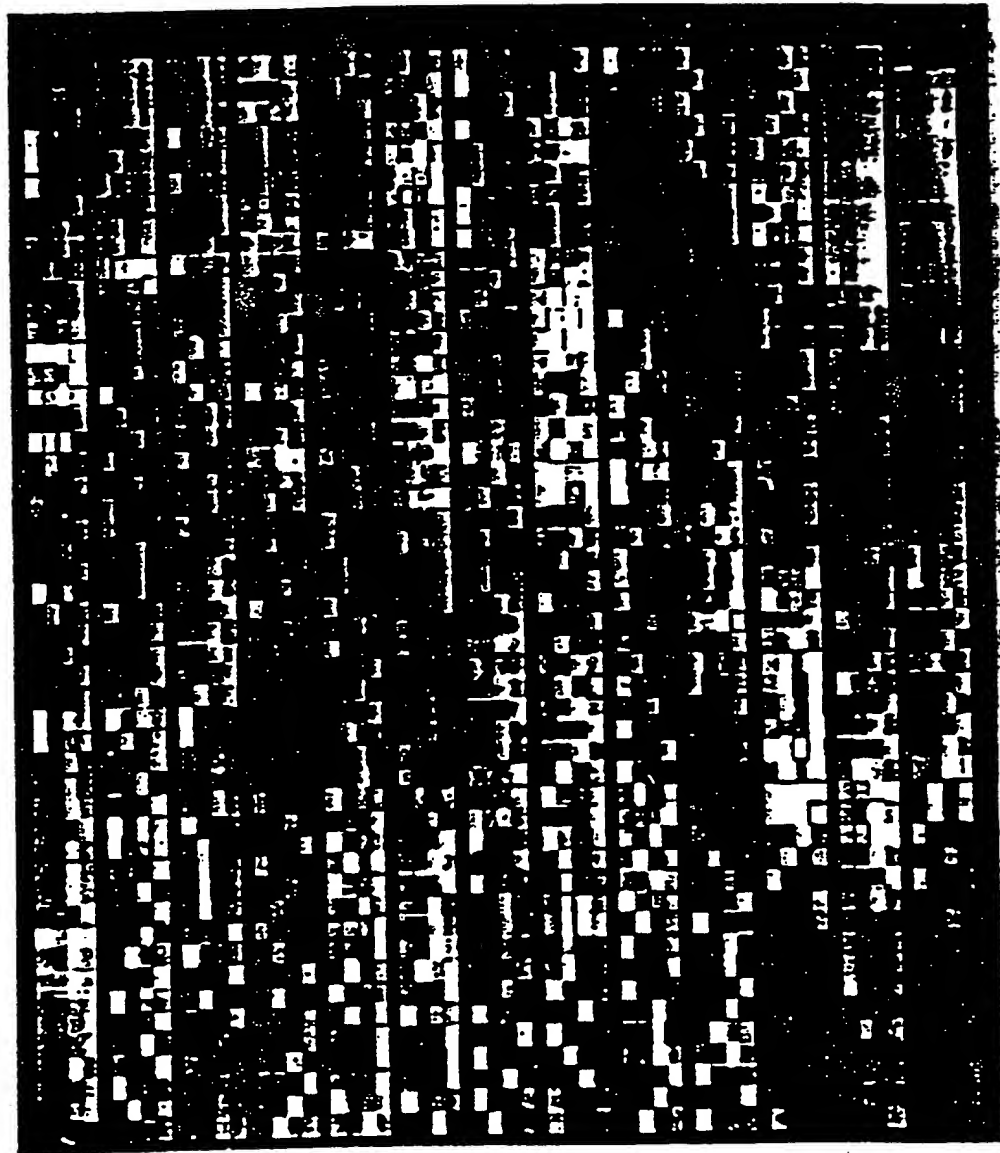
1. 50
XaacaacctaccaccccttaacagtagtacataaagccatttacX
cgtacatagcacattacagtcacaaatcccttctcgtccccaaggatgaccc
ccctcagatagggtcccttgaccaccatccctccgtgaaatcaatatccc
gcacaagagtgctactctcctcgctccgggcccataaacacttgggggtag
ctaaagtgaactglatccgacatctggttcctacttcagggtTcataaagc
ctaaatagcccacacgltccccttaaatagacatcacgattggatcacag
gtctatcacccctatlaaccactcacgggagctctccatgcatcttgggtatt
ttcgtctgggggtatgcacgcgatatgcatctgcgagacgctgggagccgga
gcaccctatgtcgcagtatctgtctttgatctcctgcctcatcttatt
tatcgcacctacgttcaatatlacaggcggaacatacttactaaagtgtgt
taattaatlaatgcttgtaggacataataataacaattgaatgtctgcac
agccActtctccacacagacatcataacaaaaaatttccaccaaaccccc
XctcccccgcttctggccacagcacttaaacacacatctTctgccaaaccccx

Fig. 43

50/57

Fig. 44

HYBRIDIZATION







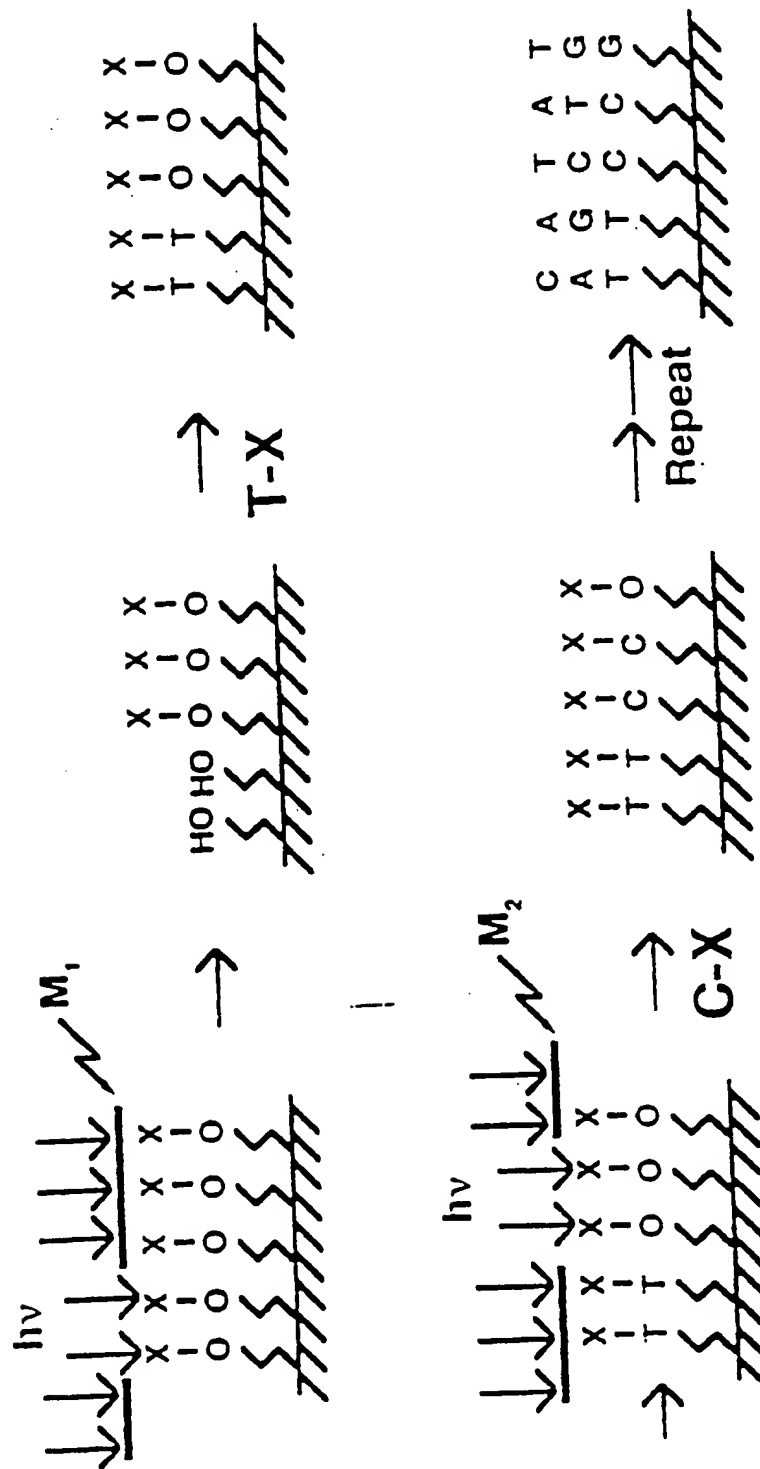
Position:	16519	152	263	344	
Change:	T->C	T->C	A->G	T->C	
Result:					T G C A

Fig. 45

52/57

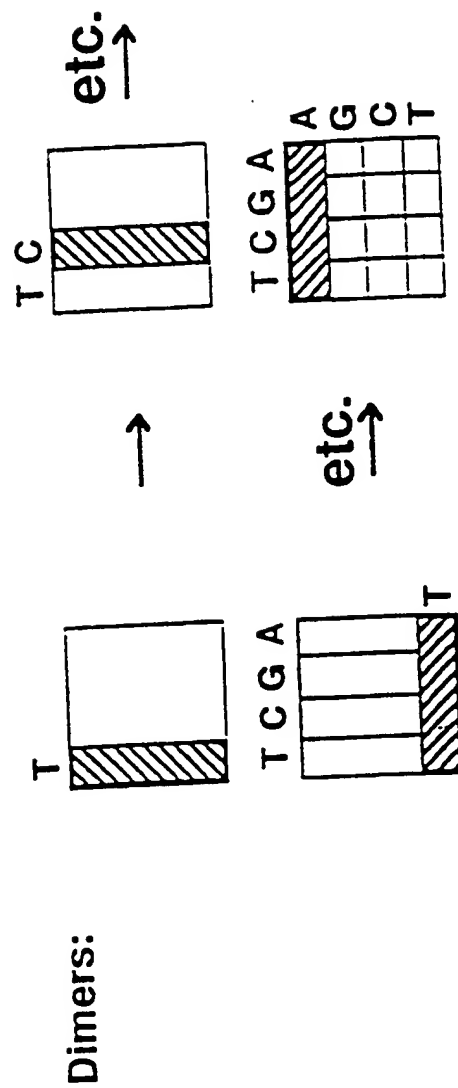
Fig. 46

Light Directed Oligonucleotide Synthesis



53/57

Nucleoside Combinatorials



in polynomial notation:
 $(T + C + A + G)^2 = \text{All Dimers}$

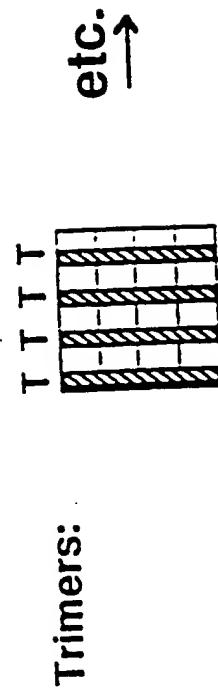
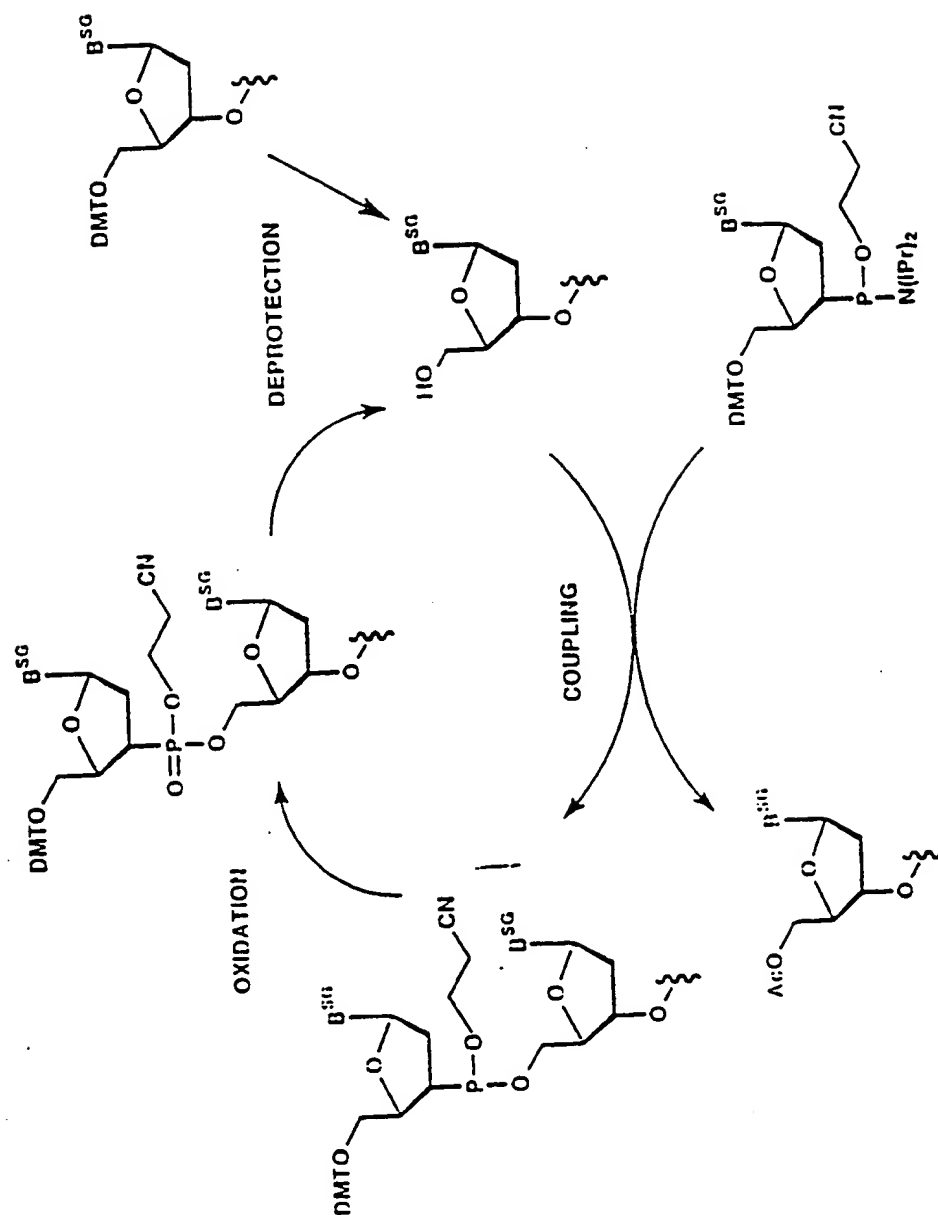


Fig. 47

54/57

Fig. 48

Solid Phase DNA Synthesis



Nucleoside Buildingblocks

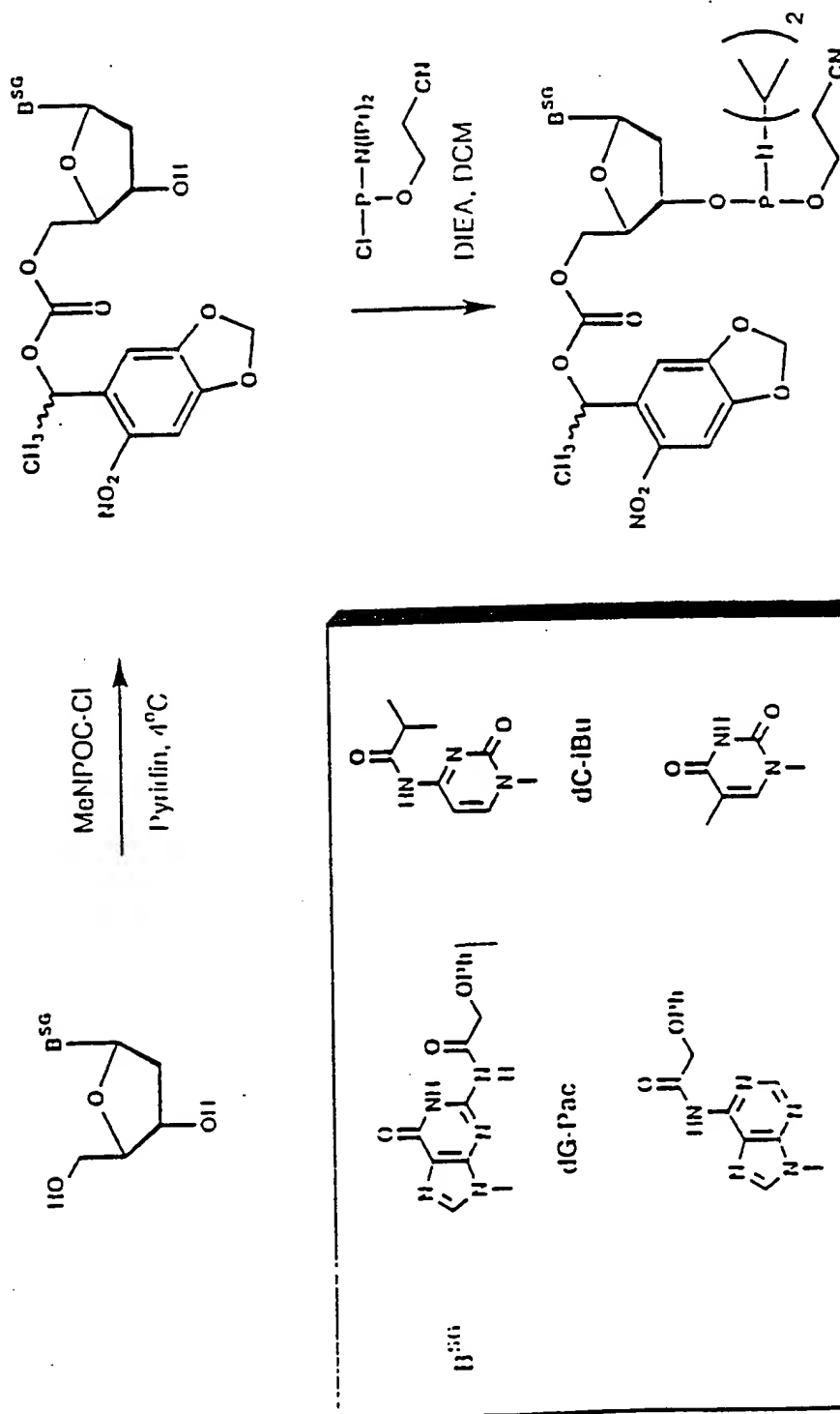
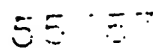
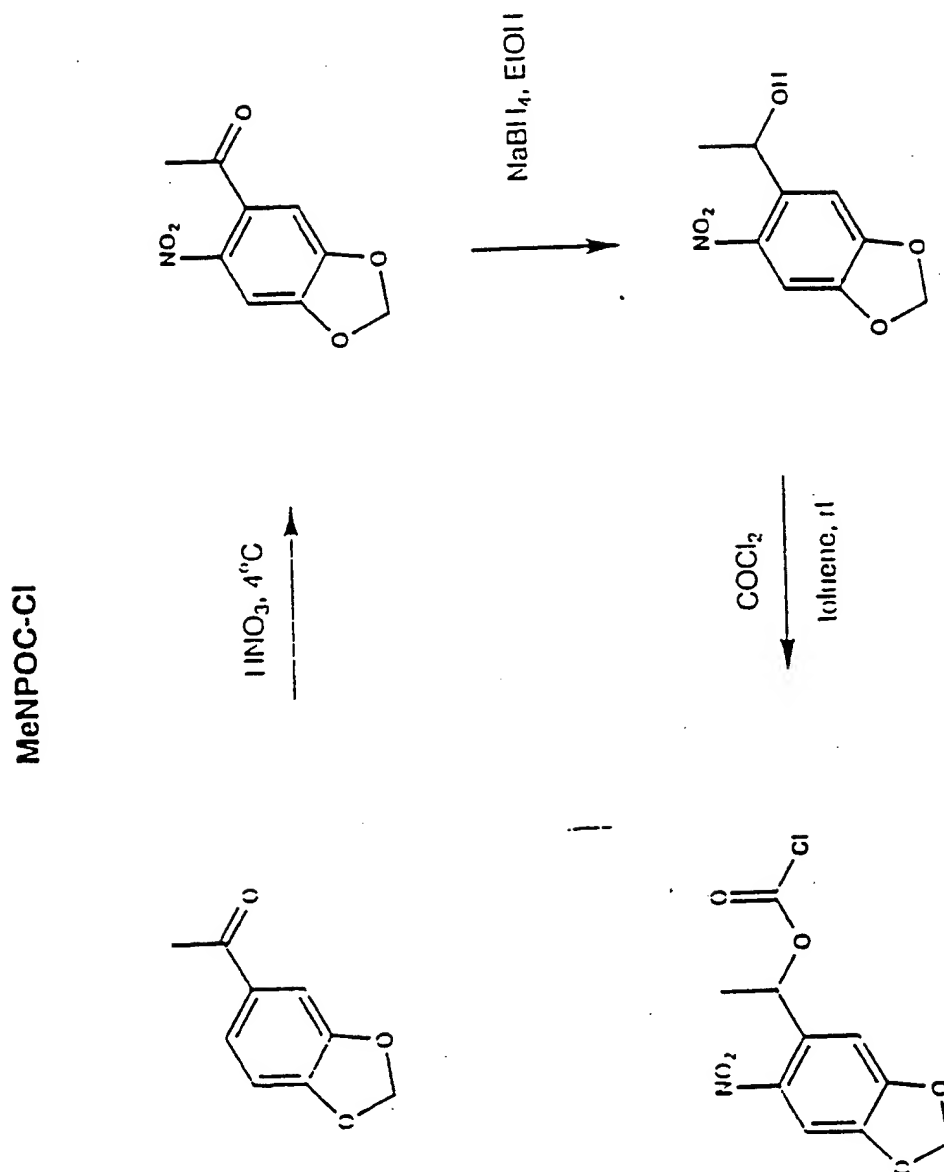


Fig. 49



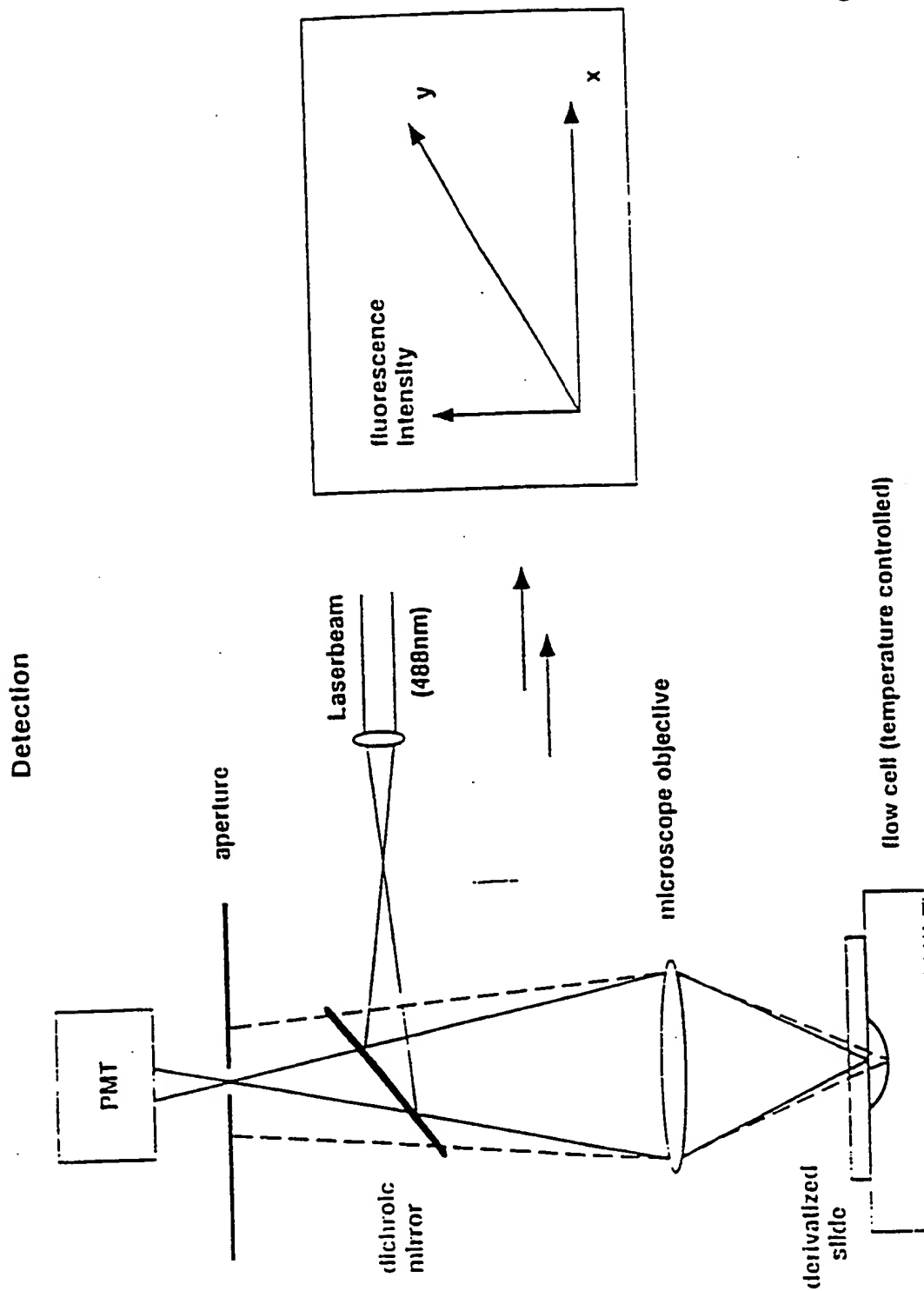
56/57

Fig. 50



57/57

Fig. 51



INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/12305

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : Please See Extra Sheet.

US CL : Please See Extra Sheet.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 810; 536/22.1, 23.1, 24.3, 24.31, 24.32, 24.33, 25.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Please See Extra Sheet.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US, A, 4,656,127 (MUNDY) 07 April 1987, see especially figure 8 and example 2 in columns 10-13.	1, 7-11, 18, 28, 29, 52, 53, 60, 61, 63 ----- 2-6, 12-17, 19- 27, 30-36, 47- 51, 54-59, 62, 64-84



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:	*T later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A document defining the general state of the art which is not considered to be of particular relevance	*X document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*E earlier document published on or after the international filing date	*Y document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*Z document member of the same patent family
*O document referring to an oral disclosure, use, exhibition or other means	
*P document published prior to the international filing date but later than the priority date claimed	

Date of the actual completion of the international search

13 FEBRUARY 1995

Date of mailing of the international search report

02 MAR 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

ARDIN MARSCHEL

Telephone No. (703) 308-0196

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US94/12305

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X — Y	WO, A, 92/10588 (FODOR ET AL.) 25 June 1992, see the entire disclosure, especially the abstract.	1-36, 47-64 ----- 65-84
Y	Nature, Volume 313, issued 24 January 1985, Ratner et al., "Complete nucleotide sequence of the AIDS virus, HTLV-III", pages 277-284, see the entire disclosure.	65-84
Y	Virology, Volume 175, issued 1990, Querat et al., "Nucleotide Sequence Analysis of SA-OMVV, a Visna-Related Ovine Lentivirus: Phylogenetic History of Lentiviruses", pages 434-447.	64-70, 72-84
Y,P	Journal of Virology, Volume 68, Number 6, issued June 1994, Luo et al., "Cellular Protein Modulates Effects of Human Immunodeficiency Virus Type 1 Rev", pages 3850-3856, see the abstract.	64-69, 71-84
Y	Cell, Volume 40, issued January 1985, Wain-Hobson et al., "Nucleotide Sequence of the AIDS Virus, LAV", pages 9-17, see the entire disclosure.	65-84
Y	Journal of Biomolecular Structure & Dynamics, Volume 11, Number 3, issued 1993, Lipshutz, "Likelihood DNA Sequencing By Hybridization", pages 637-653, see the entire disclosure and especially the abstract.	1-36, 47-84
Y	Maximum Entropy and Bayesian Methods (Paris), issued 1992, Elder, "Analysis of DNA Oligonucleotide Hybridization Data by Maximum Entropy", pages 1-10, see especially the abstract and the discussion relating to Figure 2 on page 6.	1-36, 47-84
Y	WO, A, 89/10977 (SOUTHERN) 16 November 1989, see especially the abstract and claims 1-14.	1-36, 47-84
X — Y	Genomics, Volume 13, issued 1992, Southern et al., "Analyzing and Comparing Nucleic Acid Sequences by Hybridization to Arrays of Oligonucleotides: Evaluation Using Experimental Models", pages 1008-1017, see especially the abstract and Figures 2-4 on pages 1010-1012.	1-36, 47-64 ----- 65-84
X — Y	WO, A, 93/17126 (CHETVERIN ET AL.) 02 September 1993, see especially the abstract and claims 1-197.	1-36, 47-64 ----- 65-84

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/12305

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US, A, 5,002,867 (MACEVICZ) 26 March 1991, see especially the abstract and claims 1-23.	1-36, 47-64 ----- 65-84
X --- Y	US, A, 5,202,231 (DRMANAC ET AL.) 13 April 1993, see especially the abstract and claims 1-4.	1-36, 47-64 ----- 65-84

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US94/12305**Box I Observations where certain claims were found unsearchable (Continuation of item 1 of first sheet)**

This international report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box II Observations where unity of invention is lacking (Continuation of item 2 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Please See Extra Sheet.

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-36 and 47-84

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest.
☐ No protest accompanied the payment of additional search fees.

A. CLASSIFICATION OF SUBJECT MATTER:
IPC (6):

C12Q 1/68; C07H 21/02, 21/04

A. CLASSIFICATION OF SUBJECT MATTER:
US CL :

435/6; 536/22.1, 23.1, 24.3, 24.31, 24.32

B. FIELDS SEARCHED

Electronic data bases consulted (Name of data base and where practicable terms used):

CAS, BIOSIS, WORLD PATENT INDEX, BIOTECH ABS., MEDLINE

search terms: probes, arrays, hybridization, matrix, sequencing, probe set

BOX II. OBSERVATIONS WHERE UNITY OF INVENTION WAS LACKING

This ISA found multiple inventions as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I, claims 1-36 and 47-84, drawn to arrays of oligonucleotide probes including specifically HIV directed arrays and methods of using said arrays via hybridization to target nucleic acid.

Group II, claims 37-41, drawn to methods of using arrays of pools of probes for the comparison of a target sequence with a reference sequence.

Group III, claims 42-46, drawn to pooled probes and arrays of pooled probes immobilized on a solid support.

Group IV, claims 85-96, drawn to arrays directed to reference sequences directed to the CFTR gene.

Group V, claims 97, 98, and 100-103, drawn to arrays directed to reference sequences directed to the p53 and hMLH1 genes.

Group VI, claim 99, drawn to arrays directed to reference sequences directed to the MSH2 gene.

Group VII, claims 104-108, drawn to arrays directed to reference sequences directed to sequences from the mitochondrial genome.

The inventions listed as Groups I-VII do not relate to a single inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

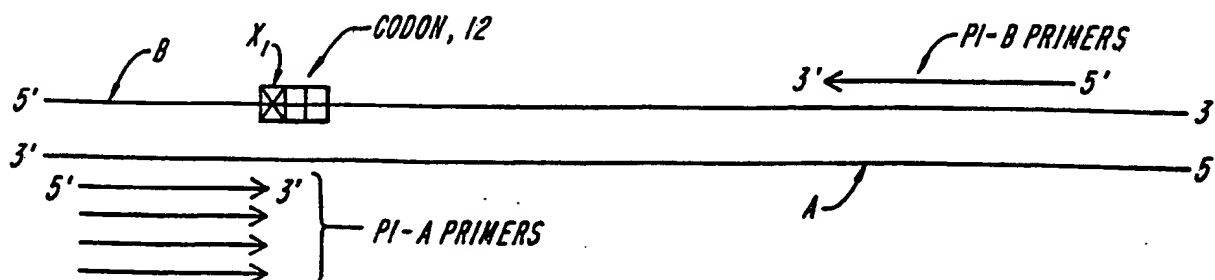
The arrays and methods of use in the invention of Group I utilize probes sets having interrogation positions therein where hybridization of the target nucleic acid to certain probes results in determining whether the target nucleic acid is the same or different from the reference sequence. Certain claims are directed specifically to HIV reference sequences. The special technical feature is deemed to be the practice of probe sets wherein specific hybridization to certain probes produces an indication of whether the interrogation position is the same or different from the reference sequence wherein the target sequence is determined by the compilation of interrogation sequence results to obtain the entire sequence. The first claimed specific reference sequence is directed to HIV. In contrast, Groups II and III cite the practice of pooled probes with variant sequences therein which are exactly complementary to each variant target sequence. The intensity of hybridization to each pool is the manner of determining the comparison between the target nucleic acid and the reference sequence. Groups II and III therefore do not determine the target sequence using the special technical features cited above but instead signal intensity using pooled probes. Therefore unity of invention is lacking between Group I and Groups II and III. Groups II and III also lack unity of invention with each other because Group II is directed to methods of using Group II is directed to the use of arrays of pooled probes whereas Group III is

directed only to a pool of probes which may have many other uses and also contain limitations therein to specific positions in probes in the pool which are not recited in Group II. Thus, Groups II and III lack unity of invention in not containing the same special technical feature for probes in a pool or pools therein. Groups IV-VII all are directed to arrays similar to that cited in Group I but are directed to completely different specific reference genes. Therefore Groups IV-VII lack unity of invention with Groups II and III for the same reasons as discussed above regarding Group I. Additionally the completely different and totally unrelated specific reference genes cited in Groups IV-VII therefore are directed to a different specific reference gene which is deemed the special technical feature of these Groups when each of Groups I and IV-VII are compared to any other Group therein. In summary the claims are not so linked by a special technical feature within the meaning of PCT Rule 13.2 so as to form a single inventive concept.



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁵: C12Q 1/68, C12P 19/34 C07H 21/04	A1	(11) International Publication Number: WO 93/22456 (43) International Publication Date: 11 November 1993 (11.11.93)
(21) International Application Number: PCT/US93/03561 (22) International Filing Date: 14 April 1993 (14.04.93) (30) Priority data: 07/874,845 27 April 1992 (27.04.92) US (71) Applicant: TRUSTEES OF DARTMOUTH COLLEGE [US/US]; 309 McNutt Building, P.O. Box 7, Hanover, NH 03755 (US). (72) Inventor: SORENSON, George, D. ; P.O. Box 176, Meriden, NH 03770 (US). (74) Agents: GEARY, William, C., III et al.; Lahive & Cockfield, 60 State Street, Boston, MA 02109 (US).		(81) Designated States: CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE). Published <i>With international search report.</i>

(54) Title: DETECTION OF GENE SEQUENCES IN BIOLOGICAL FLUIDS**(57) Abstract**

Methods are provided for detecting and quantitating gene sequences, such as mutated genes and oncogenes, in biological fluids. The fluid sample (e.g., plasma, serum, urine, etc.) is obtained, deproteinized and the DNA present in the sample is extracted. Following denaturation of the DNA, an amplification procedure, such as PCR or LCR, is conducted to amplify the mutated gene sequence.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AT	Austria	FR	France	MR	Mauritania
AU	Australia	GA	Gabon	MW	Malawi
BB	Barbados	GB	United Kingdom	NL	Netherlands
BE	Belgium	GN	Guinea	NO	Norway
BF	Burkina Faso	GR	Greece	NZ	New Zealand
BG	Bulgaria	HU	Hungary	PL	Poland
BJ	Benin	IE	Ireland	PT	Portugal
BR	Brazil	IT	Italy	RO	Romania
CA	Canada	JP	Japan	RU	Russian Federation
CF	Central African Republic	KP	Democratic People's Republic of Korea	SD	Sudan
CG	Congo	KR	Republic of Korea	SE	Sweden
CH	Switzerland	KZ	Kazakhstan	SK	Slovak Republic
CI	Côte d'Ivoire	LJ	Liechtenstein	SN	Senegal
CM	Cameroon	LK	Sri Lanka	SU	Soviet Union
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	MC	Monaco	TG	Togo
DE	Germany	MG	Madagascar	UA	Ukraine
DK	Denmark	ML	Mali	US	United States of America
ES	Spain	MN	Mongolia	VN	Viet Nam
FI	Finland				

DETECTION OF GENE SEQUENCES IN BIOLOGICAL FLUIDS

Government Support

The research leading to this invention was supported by government funding pursuant to NIH Grant No. CA 47248.

Background of the Invention

Soluble DNA is known to exist in the blood of healthy individuals at concentrations of about 5 to 10 ng/ml. It is believed that soluble DNA is present in increased levels in the blood of individuals having autoimmune diseases, particularly systemic lupus erythematosus (SLE) and other diseases including viral hepatitis, cancer and pulmonary embolism. It is not known whether circulating soluble DNA represents a specific type of DNA which is particularly prone to appear in the blood. However, studies indicate that the DNA behaves as double-stranded DNA or as a mixture of double-stranded and single-stranded DNA, and that it is likely to be composed of native DNA with single-stranded regions. Dennin, R.H., Klin. Wochenschr. 57:451-456, (1979). Steinman, C.R., J. Clin. Invest., 73:832-841, (1984). Fournie, G.J. et al., Analytical Biochem. 158:250-256, (1986). There is also evidence that in patients with SLE, the circulating DNA is enriched for human repetitive sequence (Alu) containing fragments when compared to normal human genomic DNA.

-2-

In patients with cancer, the levels of circulating soluble DNA in blood are significantly increased. Types of cancers which appear to have a high incidence of elevated DNA levels include pancreatic carcinoma, breast carcinoma, colorectal carcinoma and pulmonary carcinoma. In these forms of cancer, the levels of circulating soluble DNA in blood are usually over 50 ng/ml, and generally the mean values are more than 150 ng/ml. Leon et al., Can. Res. 37:646-650, 1977; Shapiro et al., Cancer 51:2116-2120, 1983.

Mutated oncogenes have been described in experimental and human tumors. In some instances certain mutated oncogenes are associated with particular types of tumors. Examples of these are adenocarcinomas of the pancreas, colon and lung which have approximately a 75%, 50%, and 35% incidence respectively, of Kirsten ras (K-ras) genes with mutations in positions 1 or 2 of codons 12. The most frequent mutations are changes from glycine to valine (GGT to GTT), glycine to cysteine (GGT to TGT), and glycine to aspartic acid (GGT to GAT). Other, but less common mutations of codon 12 include mutations to AGT and CGT. K-ras genes in somatic cells of such patients are not mutated.

The ability to detect sequences of mutated oncogenes or other genes in small samples of biological fluid, such as blood plasma, would provide a useful diagnostic tool. The presence of mutated K-ras gene sequences in the plasma would be indicative of the presence in the patient of a tumor

which contains mutated oncogenes. Presumably this would be a specific tumor marker since there is no other known source of mutated K-ras genes. Therefore, this evaluation may be useful in suggesting and/or confirming a diagnosis. The amount of mutated K-ras sequences in the plasma may relate to the size of the tumor, the growth rate of the tumor and/or the regression of the tumor. Therefore, serial quantitation of mutated K-ras sequences may be useful in determining changes in tumor mass. Since most human cancers have mutated oncogenes, evaluation of plasma DNA for mutated sequences may have very wide applicability and usefulness.

Summary Of The Invention

This invention recognizes that gene sequences (e.g., oncogene sequences) exist in blood, and provides a method for detecting and quantitating gene sequences such as from mutated oncogenes and other genes in biological fluids, such as blood plasma and serum. The method can be used as a diagnostic technique to detect certain cancers and other diseases which tend to increase levels of circulating soluble DNA in blood. Moreover, this method is useful in assessing the progress of treatment regimes for patients with certain cancers.

The method of the invention involves the initial steps of obtaining a sample of biological fluid (e.g., urine, blood plasma or serum, sputum, cerebral spinal fluid), then deproteinizing and extracting the DNA. The DNA is then amplified by

techniques such as the polymerase chain reaction (PCR) or the ligase chain reaction (LCR) in an allele specific manner to distinguish a normal gene sequence from a mutated gene sequence present in the sample. In one embodiment where the location of the mutation is known, the allele specific PCR amplification is performed using four pairs of oligonucleotide primers. The four primer pairs include a set of four allele specific first primers complementary to the gene sequence contiguous with the site of the mutation on the first strand. These four primers are unique with respect to each other and differ only at the 3' nucleotide which is complementary to the wild type nucleotide or to one of the three possible mutations which can occur at this known position. The four primer pairs also include a single common primer which is used in combination with each of the four unique first strand primers. The common primer is complementary to a segment of a second strand of the DNA, at some distance from the position of the first primer.

This amplification procedure amplifies a known base pair fragment which includes the mutation. Accordingly, this technique has the advantage of displaying a high level of sensitivity since it is able to detect only a few mutated DNA sequences in a background of a 10^7 -fold excess of normal DNA. The method is believed to be of much greater sensitivity than methods which detect point mutations by hybridization of a PCR product with allele specific radiolabelled probes which will not detect a mutation if the normal DNA is in more than 20-fold excess.

The above embodiment is useful where a mutation exists at a known location on the DNA. In another embodiment where the mutation is known to exist in one of two possible positions, eight pair of oligonucleotide primers may be used. The first set of four primer pairs (i.e., the four unique, allele specific primers, each of which forms a pair with a common primer) is as described above. The second set of four primer pairs comprises four allele specific primers complementary to the gene sequence contiguous with the site of the second possible mutation on the sense strand. These four primers are unique with respect to each other and differ at the terminal 3' nucleotide which is complementary to the wild type nucleotide or to one of the three possible mutations which can occur at this second known position. Each of these allele specific primers is paired with another common primer complementary to the other strand, distant from the location of the mutation.

The PCR techniques described above preferably utilize a DNA polymerase which lacks 3'exonuclease activity and therefore the ability to proofread. A preferred DNA polymerase is Thermus aquaticus DNA polymerase.

During the amplification procedure, it is usually sufficient to conduct approximately 30 cycles of amplification in a DNA thermal cycler. After an initial denaturation period of 5 minutes, each amplification cycle preferably includes a denaturation period of about 1 minute at 95°C., primer annealing for about 2 minutes at 58°C and an extension at 72°C for approximately 1 minute.

Following the amplification, aliquots of amplified DNA from the PCR can be analyzed by techniques such as electrophoresis through agarose gel using ethidium bromide staining. Improved sensitivity may be attained by using labelled primers and subsequently identifying the amplified product by detecting radioactivity or chemiluminescence on film. Labelled primers may also permit quantitation of the amplified product which may be used to determine the amount of target sequence in the original specimen.

As used herein, allele specific amplification describes a feature of the method of the invention where primers are used which are specific to a mutant allele, thus enabling amplification of the sequence to occur where there is 100% complementarity between the 3' end of the primer and the target gene sequence. Thus, allele specific amplification is advantageous in that it does not permit amplification unless there is a mutated allele. This provides an extremely sensitive detection technique.

Brief Description Of The Drawings

Figures 1A and 1B are diagrammatic representations of the amplification strategy for the detection of a mutated K-ras gene with a mutation present at a single known location of K-ras.

Figures 2A and 2B are diagrammatic representations of the amplification strategy for detection of a mutated K-ras gene with a mutation present at a second of two possible locations of K-ras.

Detailed Description of The Invention

The detection of mutated DNA, such as specific single copy genes, is potentially useful for diagnostic purposes, and/or for evaluating the extent of a disease. Normal plasma is believed to contain about 10 ng of soluble DNA per ml. The concentration of soluble DNA in blood plasma is known to increase markedly in individuals with cancer and some other diseases. The ability to detect the presence of known mutated gene sequences, such as K-ras gene sequences, which are indicative of a medical condition, is thus highly desirable.

The present invention provides a highly sensitive diagnostic method enabling the detection of such mutant alleles in biological fluid, even against a background of as much as a 10^7 -fold excess of normal DNA. The method generally involves the steps of obtaining a sample of a biological fluid containing soluble DNA, deproteinizing, extracting and denaturing the DNA, followed by amplifying the DNA in an allele specific manner, using a set of primers among which is a primer specific for the mutated allele. Through this allele specific amplification technique, only the mutant allele is amplified. Following amplification, various

techniques may be employed to detect the presence of amplified DNA and to quantify the amplified DNA. The presence of the amplified DNA represents the presence of the mutated gene, and the amount of the amplified gene present can provide an indication of the extent of a disease.

This technique is applicable to the identification in biological fluid of sequences from single copy genes, mutated at a known position on the gene. Samples of biological fluid having soluble DNA (e.g., blood plasma, serum, urine, sputum, cerebral spinal fluid) are collected and treated to deproteinize and extract the DNA. Thereafter, the DNA is denatured. The DNA is then amplified in an allele specific manner so as to amplify the gene bearing a mutation.

During deproteinization of DNA from the fluid sample, the rapid removal of protein and the virtual simultaneous deactivation of any DNase is believed to be important. DNA is deproteinized by adding to aliquots of the sample an equal volume of 20% NaCl and then boiling the mixture for about 3 to 4 minutes. Subsequently, standard techniques can be used to complete the extraction and isolation of the DNA. A preferred extraction process involves concentrating the amount of DNA in the fluid sample by techniques such as centrifugation.

The use of the 20% NaCl solution, followed by boiling, is believed to rapidly remove protein and simultaneously inactivate any DNases present. DNA

present in the plasma is believed to be in the form of nucleosomes and is thus believed to be protected from the DNases while in blood. However, once the DNA is extracted, it is susceptible to the DNases. Thus, it is important to inactivate the DNases at the same time as deproteinization to prevent the DNases from inhibiting the amplification process by reducing the amount of DNA available for amplification. Although the 20% NaCl solution is currently preferred, it is understood that other concentrations of NaCl, and other salts, may also be used.

Other techniques may also be used to extract the DNA while preventing the DNases from affecting the available DNA. Because plasma DNA is believed to be in the form of nucleosomes (mainly histones and DNA), plasma DNA could also be isolated using an antibody to histones or other nucleosomal proteins. Another approach could be to pass the plasma (or serum) over a solid support with attached antihistone antibodies which would bind with the nucleosomes. After rinsing the nucleosomes can be eluted from the antibodies as an enriched or purified fraction. Subsequently, DNA can be extracted using the above or other conventional methods.

In one embodiment, the allele specific amplification is performed through the Polymerase Chain Reaction (PCR) using primers having 3' terminal nucleotides complementary to specific point mutations of a gene for which detection is sought. PCR preferably is conducted by the method described by Saiki, "Amplification of Genomic DNA", PCR Protocols,

-10-

Eds. M.A. Innis, et al., Academic Press, San Diego (1990), pp. 13. In addition, the PCR is conducted using a thermostable DNA polymerase which lacks 3' exonuclease activity and therefore the ability to repair single base mismatches at the 3' terminal nucleotide of the DNA primer during amplification. As noted, a preferred DNA polymerase is T. aquaticus DNA polymerase. A suitable T. aquaticus DNA polymerase is commercially available from Perkin-Elmer as AmpliTaq DNA polymerase. Other useful DNA polymerases which lack 3' exonuclease activity include a Ventr (exo-), available from New England Biolabs, Inc., (purified from strains of E. coli that carry a DNA polymerase gene from the archaeobacterium Thermococcus litoralis), Hot Tub DNA polymerase derived from Thermus flavus and available from Amersham Corporation, and Tth DNA polymerase derived from Thermus thermophilus, available from Epicentre Technologies, Molecular Biology Resource Inc., or Perkin-Elmer Corp.

This method conducts the amplification using four pairs of oligonucleotide primers. A first set of four primers comprises four allele specific primers which are unique with respect to each other. The four allele specific primers are each paired with a common distant primer which anneals to the other DNA strand distant from the allele specific primer. One of the allele specific primers is complementary to the wild type allele (i.e., is allele specific to the normal allele) while the others have a mismatch at the 3' terminal nucleotide of the primer. As noted, the four unique primers are individually paired for

-11-

amplification (e.g., by PCR amplification) with a common distant primer. When the mutated allele is present, the primer pair including the allele specific primer will amplify efficiently and yield a detectable product. While the mismatched primers may anneal, the strand will not be extended during amplification.

The above primer combination is useful where a mutation is known to exist at a single position on an allele of interest. Where the mutation may exist at one of two locations, eight pair of oligonucleotide primers may be used. The first set of four pair is as described above. The second four pair of primers comprises four allele specific oligonucleotide primers complementary to the gene sequence contiguous with the site of the second possible mutation on the sense strand. These four primers differ at the terminal 3' nucleotide which is complementary to the wild type nucleotide or to one of the three possible mutations which can occur at this second known position. Each of the four allele specific primers is paired with a single common distant primer which is complementary to the antisense strand upstream of the mutation.

During a PCR amplification using the above primers, only the primer which is fully complementary to the allele which is present will anneal and extend. The primers having a non-complementary nucleotide may partially anneal, but will not extend during the amplification process. Amplification generally is allowed to proceed for a suitable number

-12-

of cycles, i.e., from about 20 to 40, and most preferably for about 30. This technique amplifies a mutation-containing fragment of the target gene with sufficient sensitivity to enable detection of the mutated target gene against a significant background of normal DNA.

The K-ras gene has point mutations which usually occur at one or two known positions in a known codon. Other oncogenes may have mutations at known but variable locations. Mutations with the K-ras gene are typically known to be associated with certain cancers such as adenocarcinomas of the lung, pancreas, and colon. Figures 1A through 2B illustrate a strategy for detecting, through PCR amplification, a mutation occurring at position 1 or 2 of the 12th codon of the K-ras oncogene. As previously noted, mutations at the first or second position of the 12th codon of K-ras are often associated with certain cancers such as adenocarcinomas of the lung, pancreas, and colon.

Referring to Figures 1A and 1B, the DNA from the patient sample is separated into two strands (A and B), which represent the sense and antisense strands. The DNA represents an oncogene having a point mutation which occurs on the same codon (i.e., codon 12) at position 1 (X_1). The allele-specific primers used to detect the mutation at position 1, include a set of four P1 sense primers (P1-A), each of which is unique with respect to the others. The four P1-A primers are complementary to a gene sequence contiguous with the site of the mutation on

-13-

strand A. The four P1-A primers preferably differ from each other only at the terminal 3' nucleotide which is complementary to the wild type nucleotide or to one of the three possible mutations which can occur at this known position. Only the P1-A primer which is fully complementary to the mutation-containing segment on the allele will anneal and extend during amplification.

A common downstream primer (P1-B), complementary to a segment of the B strand downstream with respect to the position of the P1-A primers, is used in combination with each of the P1-A primers. The P1-B primer illustrated in Figure 1 anneals to the allele and is extended during the PCR. Together, the P1-A and P1-B primers identified in Table 1 and illustrated in Figure 1B amplify a fragment of the oncogene having 161 base pairs.

Figures 2A and 2B illustrate a scheme utilizing an additional set of four unique, allele specific primers (P2-A) to detect a mutation which can occur at codon 12 of the oncogene, at position 2 (X_2). The amplification strategy illustrated in Figures 1A and 1B would be used in combination with that illustrated in Figures 2A and 2B to detect mutations at either position 1 (X_1) or position 2 (X_2) in Codon 12.

Referring to Figures 2A and 2B, a set of four unique allele specific primers (P2-A) are used to detect a mutation present at a position 2 (X_2) of codon 12. The four P2-A primers are complementary to

the genetic sequence contiguous with the site of the second possible mutation. These four primers are unique with respect to each other and preferably differ only at the terminal 3' nucleotide which is complementary to the wild type nucleotide or to one of the three possible mutations which can occur at the second known position (X₂).

A single common upstream primer (P2-B) complementary to a segment of the A strand upstream of the mutation, is used in combination with each of the unique P2-A primers. The P2-A and P2-B primers identified in Table 1 and illustrated in Figure 2B will amplify a fragment having 146 base pairs.

During the amplification procedure, the polymerase chain reaction is allowed to proceed for about 20 to 40 cycles and most preferably for 30 cycles. Following an initial denaturation period of about 5 minutes, each cycle, using the AmpliTaq DNA polymerase, typically includes about one minute of denaturation at 95° C, two minutes of primer annealing at about 58° C, and a one minute extension at 72° C. While the temperatures and cycle times noted above are currently preferred, it is noted that various modifications may be made. Indeed, the use of different DNA polymerases and/or different primers may necessitate changes in the amplification conditions. One skilled in the art will readily be able to optimize the amplification conditions.

-15-

Exemplary DNA primers which are useful in practicing the method of this invention to detect the K-ras gene, having point mutations at either the first or second position in codon 12 of the gene, are illustrated in Table 1.

TABLE 1

Primers Used to Amplify (by PCR) Position 1
and 2 Mutations at Codon 12 of K-ras Gene
(5'-3')

<u>Sequence*</u>	<u>Strand</u>	<u>P1 or P2</u>
GTGGTAGTTGGAGCTG	A	P1
GTGGTAGTTGGAGCT <u>C</u>	A	P1
GTGGTAGTTGGAGCT <u>T</u>	A	P1
GTGGTAGTTGGAGCT <u>A</u>	A	P1
CAGAGAAACCTTTATCTG	B	P1
ACTCTTGCCTACGCCAC	A	P2
ACTCTTGCCTACGCCAG	A	P2
ACTCTTGCCTACGCCAT	A	P2
ACTCTTGCCTACGCCAA	A	P2
GTACTGGTGGAGTATTT	B	P2

*Underlined bases denote mutations.

The primers illustrated in Table 1 are, of course, merely exemplary. Various modifications can be made to these primers as is understood by those having ordinary skill in the art. For example, the primers could be lengthened or shortened, however the 3' terminal nucleotides must remain the same. In

addition, some mismatches 3 to 6 nucleotides back from the 3' end may be made and would not be likely to interfere with efficacy. The common primers can also be constructed differently so as to be complementary to a different site, yielding either a longer or shorter amplified product.

In one embodiment, the length of each allele specific primer can be different, making it possible to combine multiple allele specific primers with their common distant primer in the same PCR reaction. The length of the amplified product would be indicative of which allele specific primer was being utilized with the amplification. The length of the amplified product would indicate which mutation was present in the specimen.

The primers illustrated in Table 1 and Figures 1B and 2B, and others which could be used, can be readily synthesized by one having ordinary skill in the art. For example, the preparation of similar primers has been described by Stork et al., Oncogene, 6:857-862, 1991.

Other amplification methods and strategies may also be utilized to detect gene sequences in biological fluids according to the method of the invention. For example, another approach would be to combine PCR and the ligase chain reaction (LCR). Since PCR amplifies faster than LCR and requires fewer copies of target DNA to initiate, one could use PCR as first step and then proceed to LCR. Primers such as the common primers used in the allele specific amplification described previously which span a sequence of approximately 285 base pairs in

length, more or less centered on codon 12 of K-ras, could be used to amplify this fragment, using standard PCR conditions. The amplified product (approximately a 285 base pair sequence) could then be used in a LCR or ligase detection reaction (LDR) in an allele specific manner which would indicate if a mutation was present. Another, perhaps less sensitive, approach would be to use LCR or LDR for both amplification and allele specific discrimination. The later reaction is advantageous in that it results in linear amplification. Thus the amount of amplified product is a reflection of the amount of target DNA in the original specimen and therefore permits quantitation.

LCR utilizes pairs of adjacent oligonucleotides which are complementary to the entire length of the target sequence (Barany F., PNAS 88: 189-193, 1991; Barany F., PCR Methods and Applications 1: 5-16, 1991). If the target sequence is perfectly complementary to the primers at the junction of these sequences, a DNA ligase will link the adjacent 3' and 5' terminal nucleotides forming a combined sequence. If a thermostable DNA ligase is used with thermal cycling, the combined sequence will be sequentially amplified. A single base mismatch at the junction of the oligonucleotides will preclude ligation and amplification. Thus, the process is allele specific. Another set of oligonucleotides with 3' nucleotides specific for the mutant would be used in another reaction to identify the mutant allele. A series of standard conditions could be used to detect all possible mutations at any known

site. LCR typically utilizes both strands of genomic DNA as targets for oligonucleotide hybridization with four primers, and the product is increased exponentially by repeated thermal cycling.

A variation of the reaction is the ligase detection reaction (LDR) which utilizes two adjacent oligonucleotides which are complementary to the target DNA and are similarly joined by DNA ligase (Barany F., PNAS 88:189-193, 1991). After multiple thermal cycles the product is amplified in a linear fashion. Thus the amount of the product of LDR reflects the amount of target DNA. Appropriate labeling of the primers allows detection of the amplified product in an allele specific manner, as well as quantitation of the amount of original target DNA. One advantage of this type of reaction is that it allows quantitation through automation (Nickerson et al., PNAS 87: 8923-8927, 1990).

Examples of suitable oligonucleotides for use with LCR for allele specific ligation and amplification to identify mutations at position 1 in codon 12 of the K-ras gene are illustrated below in Table 2.

TABLE 2Oligonucleotides (5'-3') for use in LCR

<u>Sequence*</u>	<u>Strand</u>	<u>P1 or P2</u>
AGCTCCA <u>A</u> CTACCACAAGTT	A1	A
GCACTCTTGCCCTACGCCACC	A2-A	A
GCACTCTTGCCCTACGCCACA	A2-B	A
GCACTCTTGCCCTACGCCAC <u>G</u>	A2-C	A
GCACTCTTGCCCTACGCCACT	A2-D	A
GGTGGCGTAGGCAAGAGTGC	B1	B
AACTTGTGGTAGTTGGAGCT	B2-A	B
AACTTGTGGTAGTTGGAGCA	B2-B	B
AACTTGTGGTAGTTGGAGC <u>C</u>	B2-C	B
AACTTGTGGTAGTTGGAGC <u>G</u>	B2-D	B

*Underlined bases denote mutations.

During an amplification procedure involving LCR four oligonucleotides are used at a time. For example, oligonucleotide A1 and, separately, each of the A2 oligonucleotides are paired on the sense strand. Also, oligonucleotide B1 and, separately, each of the B2 oligonucleotides are paired on the antisense strand. For an LCD procedure, two oligonucleotides are paired, i.e., A1 with each of the A2 oligonucleotides, for linear amplification of the normal and mutated target DNA sequence.

-20-

The method of the invention is applicable to the detection and quantitation of other oncogenes in DNA present in various biological fluids. The p53 gene is a gene for which convenient detection and quantitation could be useful because alterations in this gene are the most common genetic anomaly in human cancer, occurring in cancers of many histologic types arising from many anatomic sites. Mutations of the p53 may occur at multiple codons within the gene but 80% are localized within 4 conserved regions, or "hot spots", in exons 5, 6, 7 and 8. The most popular current method for identifying the mutations in p53 is a multistep procedure. It involves PCR amplification of exons 5-8 from genomic DNA, individually, in combination (i.e., multiplexing), or sometimes as units of more than one exon. An alternative approach is to isolate total cellular RNA, which is transcribed with reverse transcriptase. A portion of the reaction mixture is subjected directly to PCR to amplify the regions of p53 cDNA using a pair of appropriate oligonucleotides as primers. These two types of amplification are followed by single strand conformation polymorphism analysis (SSCP) which will identify amplified samples with point mutations from normal DNA by differences in mobility when electrophoresed in polyacrylamide gel. If a fragment is shown by SSCP to contain a mutation, the latter is amplified by asymmetric PCR and the sequence determined by the dideoxy-chain termination method (Murakami et al, Can. Res., 51: 3356-33612, 1991).

-21-

Further, the ligase chain reaction (LCR) may be useful with p53 since LCR is better able to evaluate multiple mutations at the same time. After determining the mutation, allele specific primers can be prepared for subsequent quantitation of the mutated gene in the patient's plasma at multiple times during the clinical course.

Preferably, the method of the invention is conducted using biological fluid samples of approximately 5ml. However, the method can also be practiced using smaller sample sizes in the event that specimen supply is limited. In such case, it may be advantageous to first amplify the DNA present in the sample using the common primers. Thereafter, amplification can proceed using the allele specific primers.

The method of this invention may be embodied in diagnostic kits. Such kits may include reagents for the isolation of DNA as well as sets of primers used in the detection method, and reagents useful in the amplification. Among the reagents useful for the kit is a DNA polymerase used to effect the amplification. A preferred polymerase is Thermus aquaticus DNA polymerase available from Perkin-Elmer as AmpliTaq DNA polymerase. For quantitation of the mutated gene sequences, the kit can also contain samples of mutated DNA for positive controls as well as tubes for quantitation by competitive PCR having the engineered sequence in known amounts.

The quantitation of the mutated K-ras sequences may be achieved using either slot blot Southern hybridization or competitive PCR. Slot blot Southern hybridization can be performed utilizing the allele specific primers as probes under relatively stringent conditions as described by Verlaan-de Vries et al., Gene 50:313-20, 1986. The total DNA extracted from 5 ml of plasma will be slot blotted with 10 fold serial dilutions, followed by hybridization to an end-labeled allele specific probe selected to be complementary to the known mutation in the particular patient's tumor DNA as determined previously by screening with the battery of allele specific primers and PCR and LCR. Positive autoradiographic signals will be graded semiquantitatively by densitometry after comparison with a standard series of diluted DNA (1-500 ng) from tumor cell cultures which have the identical mutation in codon 12 of the K-ras, prepared as slot blots in the same way.

A modified competitive PCR (Gilliland et al., Proc. Nat. Acad. Sci., USA 87:2725:79; 1990; Gilliland et al., "Competitive PCR for Quantitation of MRNA", PCR Protocols (Acad. Press), pp. 60-69, 1990) could serve as a potentially more sensitive alternative to the slot blot Southern hybridization quantitation method. In this method of quantitation, the same pair of primers are utilized to amplify two DNA templates which compete with each other during the amplification process. One template is the sequence of interest in unknown amount, i.e. mutated K-ras, and the other is an engineered deletion mutant

in known amount which, when amplified, yields a shorter product which can be distinguished from the amplified mutated K-ras sequence. Total DNA extracted from the plasma as described above will be quantitated utilizing slot blot Southern hybridization, utilizing a radiolabelled human repetitive sequence probe (BLUR8). This will allow a quantitation of total extracted plasma DNA so that the same amount can be used in each of the PCR reactions. DNA from each patient (100 ng) will be added to a PCR master mixture containing P1 or P2 allele specific primers corresponding to the particular mutation previously identified for each patient in a total volume of 400 μ l. Forty μ l of master mixture containing 10 ng of plasma DNA will be added to each of 10 tubes containing 10 μ l of competitive template ranging from 0.1 to 10 attomoles. Each reaction mixture will contain dNTPs (25 μ M final concentration including [α - 32 P]dCTP at 50 μ Ci/ml), 50 pmoles of each primer, 2mM MgCl₂, 2 units of *T. aquaticus* DNA polymerase, 1 x PCR buffer, 50 μ g/ml BSA, and water to a final volume of 40 μ l. Thirty cycles of PCR will be followed by electrophoresis of the amplified products. Bands identified by ethidium bromide will excised, counted and a ratio of K-ras sequence to deletion mutant sequence calculated. To correct for difference in molecular weight, cpm obtained for genomic K-ras bands will multiplied by 141/161 or 126/146, depending upon whether position 1 (P1) or position 2 (P2) primers are used. (The exact ratio will depend upon the length of the deletion mutant.) Data will be plotted as log ratio of deletion template DNA/K-ras DNA vs. log input deletion template DNA (Gilliland et al. 1990a, 1990b).

A modified competitive PCR could also be developed in which one primer has a modified 5' end which carries a biotin moiety and the other primer has a 5' end with a fluorescent chromophore. The amplified product can then be separated from the reaction mixture by adsorption to avidin or streptavidin attached to a solid support. The amount of product formed in the PCR can be quantitated by measuring the amount of fluorescent primer incorporated into double-stranded DNA by denaturing the immobilized DNA by alkali and thus eluting the fluorescent single stands from the solid support and measuring the fluorescence (Landgraf et al., Anal. Biochem. 182:231-235, 1991).

The competitive template preferably comprises engineered deletion mutants with a sequence comparable to the fragments of the wild type K-ras and the mutated K-ras gene amplified by the P1 and P2 series of primers described previously, except there will be an internal deletion of approximately 20 nucleotides. Therefore, the amplified products will be smaller, i.e., about 140 base pairs and 125 base pairs when the P1 primers and P2 primers are used, respectively. Thus, the same primers can be used and yet amplified products from the engineered mutants can be readily distinguished from the amplified genomic sequences.

Eight deletion mutants will be produced using the polymerase chain reaction (Higuchi et al., Nucleic Acids Res. 16:7351-67 1988); Vallette et al., Nucleic Acids Res. 17:723-33, 1989; Higuchi,

PCR Technology, Ch. 6, pp. 61-70 (Stockton Press, 1989)). The starting material will be normal genomic DNA representing the wild-type K-ras or tumor DNA from tumors which are known to have each of the possible point mutations in position one and two of codon 12. The wild-type codon 12 is GGT. The following tumor DNA can be used:

First position codon 12 mutations

G→A A549
G→T* Calu1, PR371
G→C A2182, A1698

Second position codon 12 mutations

G→A* Asp1
G→T* SW480
G→C 818-1, 181-4, 818-7

(*G→T transversions in the first or second position account for approximately 80% of the point mutations found in pulmonary carcinoma and GAT (aspartic acid) or GTT (valine) are most common in pancreatic cancer.

The deletion mutants with an approximately 20 residue deletion will be derived as previously described (Vallette et al. 1989). In summary, the P1 and P2 primers will be used in an allele specific manner with the normal DNA or with DNA from the tumor cell line with each specific mutation. Each of these would be paired for amplification with a common primer which contains the sequence of the common

-26-

primer normally used with either the P1 and P2 allele specific primers, i.e., "P1-B" or "P2-B" at the 5' end with an attached series of residues representing sequences starting approximately 20 bases downstream, thus spanning the deleted area (common deletion primer 1 and 2, CD1 and CD2). The precise location and therefore sequence of the 3' portion of the primer will be determined after analysis of the sequence of the ras gene in this region with OLIGO (NBI, Plymouth, MN), a computer program which facilitates the selection of optimal primers. The exact length of the resultant amplified product is not critical, so the best possible primer which will produce a deletion of 20-25 residues will be selected. For example, with P2 primers the allele specific primer for the wild-type will be 5' ACTCTTGCCCTACGCCAC 3' complementary to residues 35 to 51 in the coding sequence. To effect a deletion of approximately 20 residues in the complementary strand, the common upstream primer to be used with the wild-type and the three allele specific primers for mutations in position two of codon 12 will be 40 residues long (CD2) complementary to residues -95 to -78 (the currently preferred common upstream primer for use with P2 allele specific primers and residues at approximately -58 to -25). The amplified shorter product will be size-separated by gel electrophoresis and purified by Prep-a-Gene (Biorad). DNA concentrations will be determined by the ethidium bromide staining with comparison to dilutions of DNA of known concentration. This approach will be repeated eight times, using the four P1 primers and common primer (CD1) constructed as above, and four

times with the four P2 primers and common primer (CD2). These deletion mutants will be amplified, using the same allele specific primers used to amplify the genomic DNA. Therefore, they can be used subsequently in known serial dilutions in a competitive PCR, as outlined above.

The invention is further illustrated by the following non-limiting examples.

Example 1

Blood was collected in 13 x 75 mm vacutainer tubes containing 0.05 ml of 15% K₃EDTA. The tubes were immediately centrifuged at 4°C for 30 minutes at 1000 g, the plasma was removed and recentrifuged at 4°C for another 30 minutes at 1000 g.

The plasma was stored at -70°C. Next, DNA was deproteinized by adding an equal volume of 20% NaCl to 5 ml aliquots of plasma which were then boiled for 3 to 4 minutes. After cooling, the samples were centrifuged at 3000 rpm for 30 minutes. The supernatant was removed and dialysed against three changes of 10 mM Tris-HCl (pH 7.5)/1 mM EDTA (pH 8.0) ("TE") for 18 to 24 hours at 4°C. The DNA was extracted once with two volumes of phenol, 2x1 volume phenol:chloroform: isoamyl alcohol (25:24:1) and 2x1 volume chloroform: isoamyl alcohol (24:1). DNA was subsequently precipitated with NaCl at 0.3M, 20µg/ml glycogen as a carrier and 2.5 volumes of 100% ethanol at minus 20°C for 24 hours. DNA was recovered by centrifugation in an Eppendorf Centrifuge at 4°C for 30 minutes. The DNA was then resuspended in a TE buffer. The DNA extracted and prepared in the above manner was then able to be amplified.

Example 2

An allele specific amplification of DNA obtained and prepared according to example 1 was conducted by PCR as follows to detect the K-ras gene in the DNA having a mutation at position 1 or 2 of the codon 12 of the K-ras gene. In each of eight reaction tubes was added DNA extracted from 0.5 ml of plasma in total volume of 40 μ l containing 67 mM Tris-HCl (pH 8.8), 10 mM β -mercaptoethanol, 16.6 μ M ammonium sulfate, 6.7 μ M EDTA, 2.0mM, MgCl₂, 50 μ g/ml BSA, 25 μ M dNTP. Also, 50 pmoles of each of the primers identified in Table 1 was included, together with 3 units of Thermus aquaticus DNA polymerase (available from Perkin-Elmer as AmpliTaq). PCR was conducted with an initial denaturation at 95°C for 5 minutes, followed by 30 cycles of PCR amplification in a DNA thermal cycler (Cetus; Perkin-Elmer Corp. Norwalk, Connecticut). Each amplification cycle includes a 1 minute denaturation at 95°C, a 2 minute primer annealing period at 58°C, and a 1 minute extension period at 72°C.

Following the completion of amplification, 10-15 μ l of each of the PCR reaction products is analyzed by electrophoresis in a 2% agarose gel/1X TAE-0.5 μ g/ml EtBr. The electrophoresis uses an applied voltage of 100 volts for 90 minutes. Photographs of the samples are then taken using ultraviolet light under standard conditions.

It is understood that various modifications can be made to the present invention without departing from the scope of the claimed invention.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

- (i) APPLICANT: Sorenson, George D.
- (ii) TITLE OF INVENTION: Detection of
Gene Sequences
In Biological
Fluids
- (iii) NUMBER OF SEQUENCES: 20
- (iv) CORRESPONDENCE ADDRESS:
 - (A) ADDRESSEE: Lahive & Cockfield
 - (B) STREET: 60 State Street
 - (C) CITY: Boston
 - (D) STATE: Massachusetts
 - (E) COUNTRY: U.S.A.
 - (F) ZIP: 02109
- (v) COMPUTER READABLE FORM:
 - (A) MEDIUM TYPE: Floppy Disk
 - (B) COMPUTER: IBM PC compatible
 - (C) OPERATING SYSTEM: PC-DOS/MS-DOS
 - (D) SOFTWARE: ASCII Text
- (vi) CURRENT APPLICATION DATA:
 - (A) APPLICATION NUMBER:
 - (B) FILING DATE: 27 APR - 1992
 - (C) CLASSIFICATION
- (viii) ATTORNEY/AGENT INFORMATION:
 - (A) NAME: William C. Geary III
 - (B) REGISTRATION NUMBER: 31,357
 - (C) REFERENCE/DOCKET NUMBER: DCI-037
- (ix) TELECOMMUNICATION INFORMATION:
 - (A) TELEPHONE (617) 227-7400
 - (B) TELEFAX: (617) 227-5941
- (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1

-30-

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:1:

GTGGTAGTTG GAGCTG

16

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:2:

GTGGTAGTTG GAGCTC

16

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 16 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GTGGTAGTTG GAGCTT

16

-31-

(2) INFORMATION FOR SEQ ID NO:4:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 16 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA
- (iii) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GTGCTAGTTG GAGCTA

16

(2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 18 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA
- (iii) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CAGAGAAACC TTTATCTG

18

(2) INFORMATION FOR SEQ ID NO:6:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 17 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: single
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: DNA
- (iii) SEQUENCE DESCRIPTION: SEQ ID NO:6:

ACTCTTGCCT ACGCCAC

17

-32-

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 17 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:7:

ACTCTTGCCT ACGCCAG

17

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 17 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:8:

ACTCTTGCCT ACGCCAA

17

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 17 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:9:

ACTCTTGCCT ACGCCAT

17

-33-

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 17 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GTACTGGTGG AGTATTT

17

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:11:

AGCTCCAACT ACCACAAGTT

20

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:12:

GCACTCTTGC CTACGCCACC

20

-34-

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GCACTCTTGC CTACGCCACA

20

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:14:

GCACTCTTGC CTACGCCACG

20

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GCACTCTTGC CTACGCCACT

20

-35-

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GGTGGCGTAG GCAAGAGTGC

20

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:17:

AACTTGTGGT AGTTGGAGCT

20

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:18:

AACTTGTGGT AGTTGGAGCA

20

-36-

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:19:

AACTTGTGGT AGTTGGAGCC

20

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 20 base pairs

(B) TYPE: nucleic acid

(C) STRANDEDNESS: single

(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA

(iii) SEQUENCE DESCRIPTION: SEQ ID NO:20:

AACTTGTGGT AGTTGGAGCG

20

Claims:

1. A method of detecting a mutant allele, comprising the steps of:

providing a sample of a biological fluid containing soluble DNA, including a mutant allele of interest;

extracting the DNA from the sample;

denaturing the DNA to free first and second strands of the DNA;

amplifying the mutant allele of interest in an allele specific manner using at least a first set of four allele specific oligonucleotide primers having one primer complementary to a mutation-containing segment on a first strand of the DNA and a first common primer for pairing during amplification to each allele specific primer, the common primer being complementary to a segment of a second strand of the DNA distant with respect to the position of the first primer; and

detecting the presence of the mutant allele of interest.

2. The method of claim 1 further comprising the step of removing protein from the sample and inactivating any DNase within the sample before the step of extracting the DNA.

-38-

3. The method of claim 2, wherein the mutant allele is amplified in an allele specific manner using the polymerase chain reaction (PCR).

4. The method of claim 3, wherein following the amplification step, the step of detecting the presence of the mutant allele of interest comprises performing an allele specific ligase chain reaction (LCR) or a ligase detection reaction (LDR) using the amplified product of PCR.

5. The method of claim 2 wherein protein is removed and DNases are inactivated by adding a salt solution to the sample and subsequently boiling the sample.

6. The method of claim 2 wherein the biological fluid is selected from the group consisting of whole blood, serum, plasma, urine, sputum, and cerebral spinal fluid.

7. The method of claim 2 wherein the mutant allele comprises a gene sequence having a point mutation at a known location.

8. The method of claim 7 wherein the first DNA strand is the sense strand and the second DNA strand is the antisense strand.

9. The method of claim 2 wherein the step of amplifying the mutant allele with the PCR is conducted using a DNA polymerase which lacks the 3' exonuclease activity and therefore the ability to repair single nucleotide mismatches at the 3' end of the primer.

-39-

10. The method of claim 9 wherein the DNA polymerase is a Thermus aquaticus DNA polymerase.

11. The method of claim 9 wherein the first set of allele specific oligonucleotide primers comprises:

four sense primers, one of which has a 3' terminal nucleotide complementary to a point mutation of the sense strand, and the remaining three of which are complementary to the wild type sequence for the segment to be amplified and to sequences having the remaining two possible mutations at the mutated point of the sense strand; and

a common antisense primer complementary to a segment of the antisense strand distant from the location on the sense strand at which the sense primers will anneal, the common antisense primer being paired with each of the sense primers during amplification.

12. The method of claim 11 wherein the 3' terminal nucleotide of the complementary sense primer anneals with the mutated nucleotide of the sense strand.

13. The method of claim 3 wherein the mutant allele comprises a gene sequence having a point mutation at one of two known locations.

-40-

14. The method of claim 13 wherein the step of amplifying the mutant allele through the PCR further comprises the use of a second set of four allele specific oligonucleotide primers, in conjunction with the first set, wherein the second set of allele specific oligonucleotide primers comprises:

four sense primers, one of which has a 3' terminal nucleotide complementary to a point mutation of the sense strand, and the remaining three of which are complementary to the wild type sequence for the segment to be amplified and sequences having the remaining two possible mutations at the mutated point of the sense strand; and

a common antisense primer complementary to a segment of the antisense strand distant from the location on the sense strand at which the sense primers will anneal, the common antisense primer being paired with each of the sense primers during amplification.

15. The method of claim 14 wherein the 3' terminal nucleotide of the complementary sense primer anneals with the mutated nucleotide of the sense strand.

16. The method of claim 15 wherein the mutant allele to be detected is the K-ras gene sequence having a mutation at position 1 or 2 in the twelfth codon.

-41-

17. The method of claim 16 wherein the first set of allele specific oligonucleotide primers comprises sense primers having the following sequences

5'GTGGTAGTTGGAGCTG 3' (wild type)
5'GTGGTAGTTGGAGCTC 3'
5'GTGGTAGTTGGAGCTT 3'
5'GTGGTAGTTGGAGCTA 3'

and the common antisense primer having the following sequence

5'CAGAGAAACCTTTATCTG 3'.

18. The method of claim 14 wherein the second set of allele specific oligonucleotide primers comprises sense primers having the following sequences

5'ACTCTTGCCTACGCCAC 3' (wild type)
5'ACTCTTGCCTACGCCAG 3'
5'ACTCTTGCCTACGCCAT 3'
5'ACTCTTGCCTACGCCAA 3'

and the common antisense primer having the following sequence

5'GTACTGGTGGAGTATTT 3'.

19. The method of claim 2 wherein the step of detecting the presence of amplified DNA is conducted by gel electrophoresis in 1-5% agarose gel.

20. The method of claim 2 wherein the biological fluid is selected from the group consisting of whole blood, serum, plasma, urine, sputum, and cerebral spinal fluid.

21. A diagnostic kit for detecting the presence of a mutated K-ras gene sequence in biological fluid, wherein the mutation is present in the twelfth codon at position 1, comprising:

reagents to facilitate the deproteinization and isolation of DNA;
reagents to facilitate amplification by PCR;

a heat stable DNA polymerase; and
a first set of allele specific oligonucleotide sense primers having the following sequences

5'GTGGTAGTTGGAGCTG 3'

5'GTGGTAGTTGGAGCTC 3'

5'GTGGTAGTTGGAGCTT 3'

5'GTGGTAGTTGGAGCTA 3'

and a first common antisense primer having the following sequence

5'CAGAGAAACCTTTATCTG '3

22. The diagnostic kit of claim 21 further comprising

a second set of allele specific oligonucleotide sense primers having the following sequences

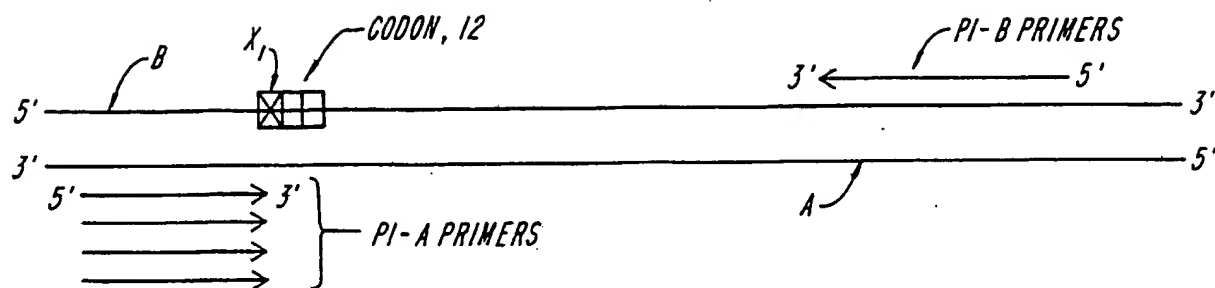
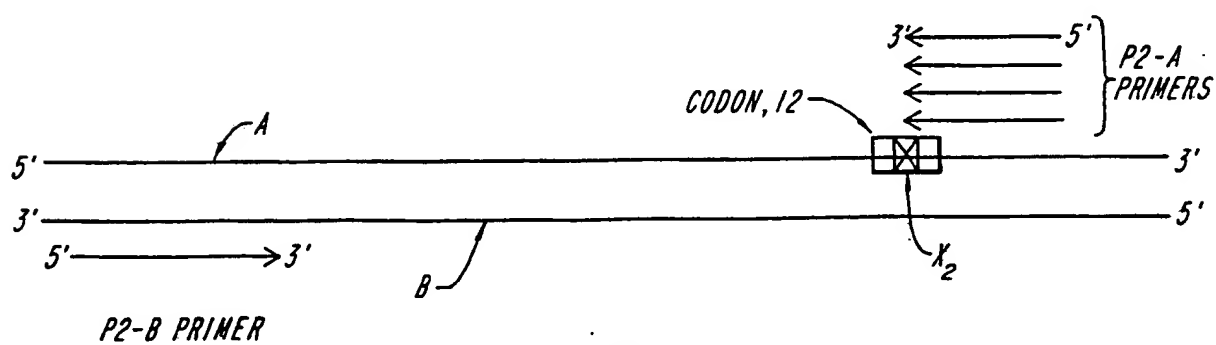
-43-

5'ACTCTTGCCTACGCCAC 3'
5'ACTCTTGCCTACGCCAG 3'
5'ACTCTTGCCTACGCCAT 3'
5'ACTCTTGCCTACGCCAA 3'

and a second common antisense primer having
the following sequence

5'GTACTGGTGGAGTATTT 3'

wherein the second set of allele specific
oligonucleotide primers and the second common primer
are useful in detecting in biological fluid the
presence of a mutated K-ras gene sequence in the
twelfth codon at position 2.

**FIG. 1A****FIG. 2A**

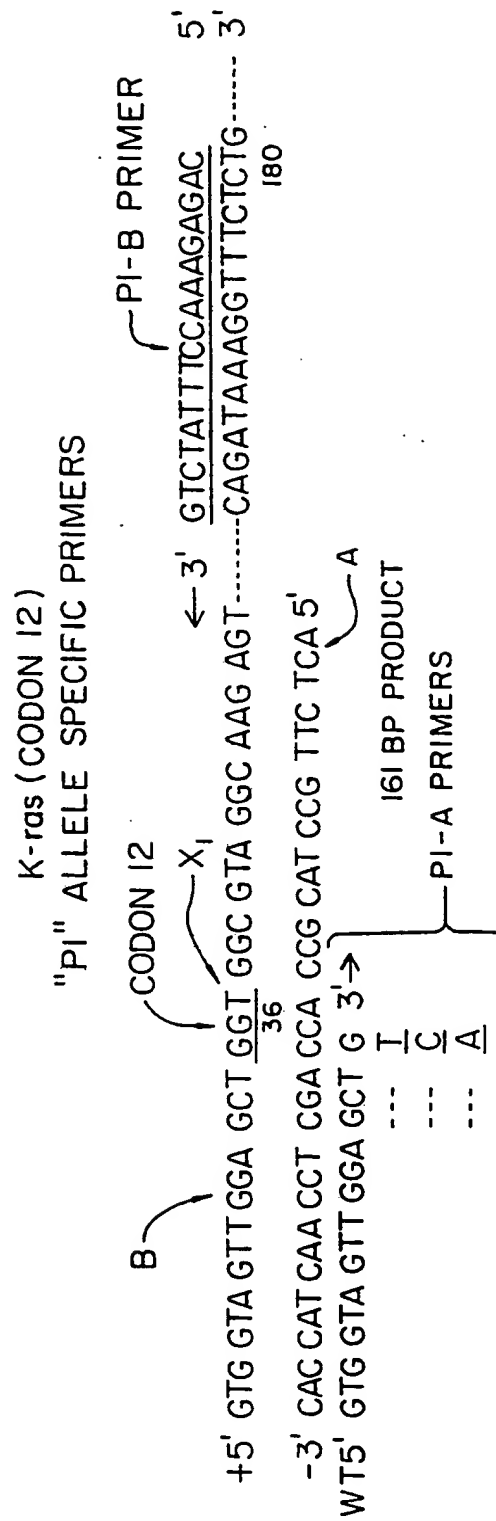


FIG. 1B

K-ras (CODON 12)
"P2" ALLELE SPECIFIC PRIMERS

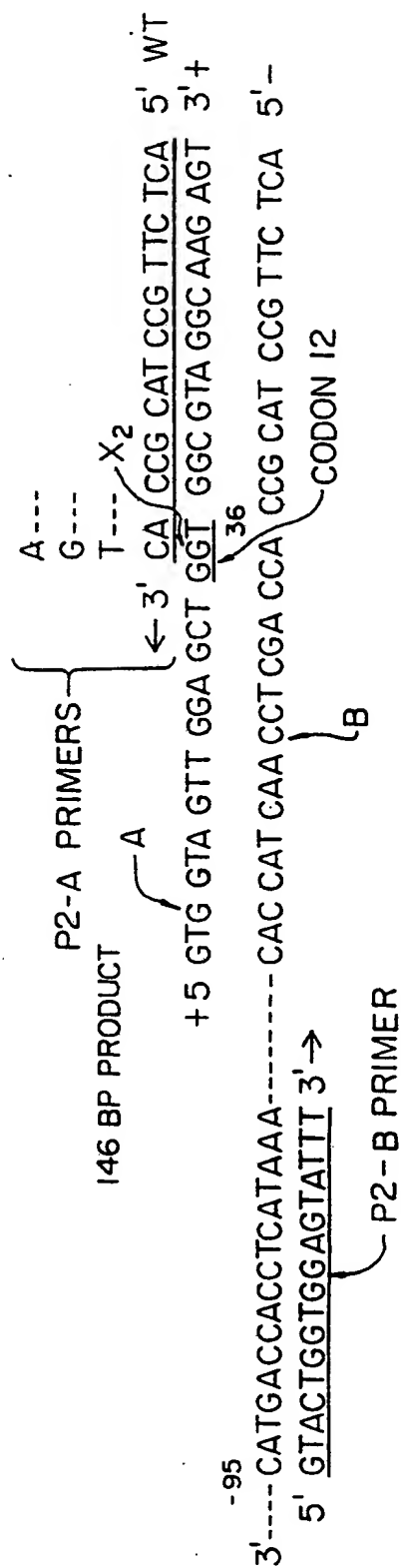


FIG. 2B

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US93/03561

A. CLASSIFICATION OF SUBJECT MATTER

IPC(5) : C12Q 1/68; C12P 19/34; C07H 21/04

US CL : 435/6, 91; 536/24.33

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6, 91; 536/24.33; 935/78

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, BIOSIS, EMBASE, MEDLINE, SCISEARCH, EMBL, GENBANK

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ONCOGENE, Volume 6, issued May 1991 by MacMillan Press Ltd., P. Stork et al., "Detection of K-ras mutations in pancreatic and hepatic neoplasms by non-isotopic mismatched polymerase chain reaction", pages 857-862, see entire document.	1-22
Y	WO, 89/00206 (BALAZS ET AL) 12 NOVEMBER 1989, see entire document.	1-20
Y	US, A, 4,988,617 (LANDEGREN ET AL) 29 JANUARY 1991, see entire document.	1-22
A.P	US, A, 5,137,806 (LeMAISTRE ET AL) 11 AUGUST 1992, see entire document.	1-22



Further documents are listed in the continuation of Box C.



See patent family annex.

*

Special categories of cited documents:

"T"

later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"A"

document defining the general state of the art which is not considered to be part of particular relevance

"X"

document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"E"

earlier document published on or after the international filing date

"L"

document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"Y"

document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"O"

document referring to an oral disclosure, use, exhibition or other means

"P"

document published prior to the international filing date but later than the priority date claimed

"&"

document member of the same patent family

Date of the actual completion of the international search

08 JUNE 1993

Date of mailing of the international search report

JUL 13 1993

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No. NOT APPLICABLE

Authorized officer

PAUL B. TRAN, PH.D.

Telephone No. (703) 308-0196

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau



B4

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C07K 14/00		A2	(11) International Publication Number: WO 98/14470
			(43) International Publication Date: 9 April 1998 (09.04.98)
(21) International Application Number: PCT/US97/18032			(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).
(22) International Filing Date: 3 October 1997 (03.10.97)			
(30) Priority Data: 08/725,885 4 October 1996 (04.10.96) US			
(71) Applicant: GENETICS INSTITUTE, INC. [US/US]; 87 CambridgePark Drive, Cambridge, MA 02140 (US).			
(72) Inventors: JACOBS, Kenneth; 151 Beaumont Avenue, Newton, MA 02160 (US). MCCOY, John, M.; 56 Howard Street, Reading, MA 01867 (US). LAVALLIE, Edward, R.; 90 Green Meadow Drive, Tewksbury, MA 01876 (US). RACIE, Lisa, A.; 124 School Street, Acton, MA 01720 (US). MERBERG, David; 2 Orchard Drive, Acton, MA 01720 (US). TREACY, Maurice; 93 Walcott Road, Chestnut Hill, MA 02167 (US). SPAULDING, Vikki; 11 Meadowbank Road, Billerica, MA 01821 (US).			
(74) Agent: SPRUNGER, Suzanne, A.; Genetics Institute, Inc., 87 CambridgePark Drive, Cambridge, MA 02140 (US).			

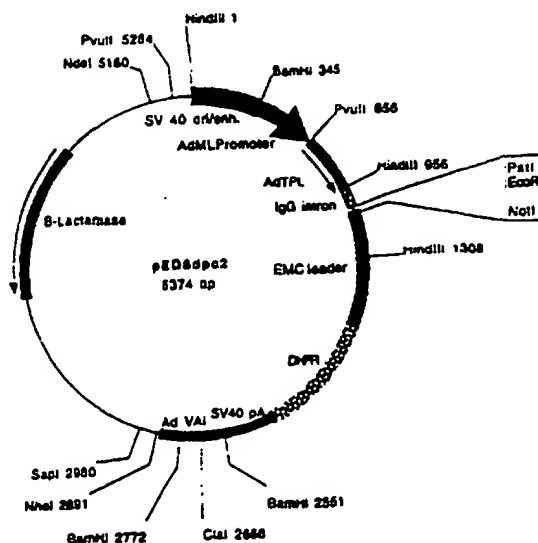
Published

Without international search report and to be republished upon receipt of that report.

(54) Title: SECRETED PROTEINS

(57) Abstract

Novel proteins are disclosed.



Plasmid name: pED8dpc2
Plasmid size: 5374 bp

Comments/References: pED8dpc2 is derived from pED8dpc1 by insertion of a new polylinker to facilitate cDNA cloning. EST cDNAs are cloned between EcoRI and NotI. pED vectors are described in Kaufman et al. (1991), MAR 19: 4485-4490.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Larvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

SECRETED PROTEINS

5

FIELD OF THE INVENTION

The present invention provides novel proteins , along with therapeutic, diagnostic and research utilities for these proteins.

BACKGROUND OF THE INVENTION

10

Technology aimed at the discovery of protein factors (including e.g., cytokines, such as lymphokines, interferons, CSFs and interleukins) has matured rapidly over the past decade. The now routine hybridization cloning and expression cloning techniques clone novel polynucleotides "directly" in the sense that they rely on information directly related to the discovered protein (i.e., partial DNA/amino acid sequence of the protein in the case of hybridization cloning; activity of the protein in the case of expression cloning). More recent "indirect" cloning techniques such as signal sequence cloning, which isolates DNA sequences based on the presence of a now well-recognized secretory leader sequence motif, as well as various PCR-based or low stringency hybridization cloning techniques, have advanced the state of the art by making available large numbers of DNA/amino acid sequences for proteins that are known to have biological activity by virtue of their secreted nature in the case of leader sequence cloning, or by virtue of the cell or tissue source in the case of PCR-based techniques. It is to these proteins that the present invention is directed.

15

20

SUMMARY OF THE INVENTION

25

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

(a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1;

30

(b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:1 from nucleotide 28 to nucleotide 276;

(c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AE402_1i deposited under accession number ATCC 98190;

35

(d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AE402_1i deposited under accession number ATCC 98190;

(e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AE402_1i deposited under accession number ATCC 98190;

5 (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AE402_1i deposited under accession number ATCC 98190;

(g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:2;

(h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:2 having biological activity;

10 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and

(j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:1
15 from nucleotide 28 to nucleotide 276; the nucleotide sequence of the full length protein coding sequence of clone AE402_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AE402_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AE402_1i deposited
20 under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 25 (a) the amino acid sequence of SEQ ID NO:2;
(b) fragments of the amino acid sequence of SEQ ID NO:2; and
(c) the amino acid sequence encoded by the cDNA insert of clone AE402_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:2.

30 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:4;
(b) a polynucleotide comprising the nucleotide sequence of SEQ ID
35 NO:4 from nucleotide 61 to nucleotide 513;

- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:4 from nucleotide 322 to nucleotide 513;
- (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AE610_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AE610_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AE610_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AE610_1i deposited under accession number ATCC 98190;
- (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:5;
- (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:5 having biological activity;
- (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and
- (k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:4 from nucleotide 61 to nucleotide 513; the nucleotide sequence of SEQ ID NO:4 from nucleotide 322 to nucleotide 513; the nucleotide sequence of the full length protein coding sequence of clone AE610_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AE610_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AE610_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:5;
- (b) fragments of the amino acid sequence of SEQ ID NO:5; and
- (c) the amino acid sequence encoded by the cDNA insert of clone AE610_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:5.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- 5 (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:7;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:7 from nucleotide 20 to nucleotide 523;
- (c) a polynucleotide comprising the nucleotide sequence of the full length
10 protein coding sequence of clone AH106_1i deposited under accession number ATCC 98190;
- (d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AH106_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide comprising the nucleotide sequence of the mature
15 protein coding sequence of clone AH106_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AH106_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding a protein comprising the amino acid
20 sequence of SEQ ID NO:8;
- (h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:8 having biological activity;
- (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-
(f) above; and
- 25 (j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:7 from nucleotide 20 to nucleotide 523; the nucleotide sequence of the full length protein coding sequence of clone AH106_1i deposited under accession number ATCC 98190; or the
30 nucleotide sequence of the mature protein coding sequence of clone AH106_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AH106_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:8;
- 5 (b) fragments of the amino acid sequence of SEQ ID NO:8; and
- (c) the amino acid sequence encoded by the cDNA insert of clone AH106_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:8.

10 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:9;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:9
15 from nucleotide 130 to nucleotide 309;
- (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AH196_1i deposited under accession number ATCC 98190;
- (d) a polynucleotide encoding the full length protein encoded by the
20 cDNA insert of clone AH196_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AH196_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide encoding the mature protein encoded by the cDNA
25 insert of clone AH196_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:10;
- (h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:10 having biological activity;
- 30 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and
- (j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:9
35 from nucleotide 130 to nucleotide 309; the nucleotide sequence of the full length protein

coding sequence of clone AH196_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AH196_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AH196_1i
5 deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:10;
- 10 (b) fragments of the amino acid sequence of SEQ ID NO:10; and
- (c) the amino acid sequence encoded by the cDNA insert of clone AH196_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:10.

15 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:12;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID
20 NO:12 from nucleotide 69 to nucleotide 467;
- (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AI6_1i deposited under accession number ATCC 98190;
- (d) a polynucleotide encoding the full length protein encoded by the
25 cDNA insert of clone AI6_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AI6_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide encoding the mature protein encoded by the cDNA
30 insert of clone AI6_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:13;
- (h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:13 having biological activity;

- (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and
- (j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

5 Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:12 from nucleotide 69 to nucleotide 467; the nucleotide sequence of the full length protein coding sequence of clone AI6_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AI6_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full
10 length or mature protein encoded by the cDNA insert of clone AI6_1i deposited under accession number ATCC 98190. In yet other preferred embodiments, such polynucleotide encodes a protein comprising the amino acid sequence of SEQ ID NO:13 from amino acid 69 to amino acid 133.

In other embodiments, the present invention provides a composition comprising a
15 protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:13;
- (b) the amino acid sequence of SEQ ID NO:13 from amino acid 69 to amino acid 133;
- 20 (c) fragments of the amino acid sequence of SEQ ID NO:13; and
- (d) the amino acid sequence encoded by the cDNA insert of clone AI6_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:13 or the amino acid sequence of SEQ ID
25 NO:13 from amino acid 69 to amino acid 133.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:16;
- 30 (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:16 from nucleotide 55 to nucleotide 337;
- (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AJ13_1i deposited under accession number ATCC 98190;

(d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AJ13_1i deposited under accession number ATCC 98190;

(e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AJ13_1i deposited under accession number ATCC 98190;

(f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AJ13_1i deposited under accession number ATCC 98190;

(g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:17;

(h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:17 having biological activity;

(i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and

(j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:16 from nucleotide 55 to nucleotide 337; the nucleotide sequence of the full length protein coding sequence of clone AJ13_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AJ13_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AJ13_1i deposited under accession number ATCC 98190. In yet other preferred embodiments, such polynucleotide encodes a protein comprising the amino acid sequence of SEQ ID NO:17 from amino acid 12 to amino acid 94.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

(a) the amino acid sequence of SEQ ID NO:17;

(b) the amino acid sequence of SEQ ID NO:17 from amino acid 12 to amino acid 94;

(c) fragments of the amino acid sequence of SEQ ID NO:17; and

(d) the amino acid sequence encoded by the cDNA insert of clone AJ13_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:17 or the amino acid sequence of SEQ ID NO:17 from amino acid 12 to amino acid 94.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:19;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:19 from nucleotide 33 to nucleotide 422;
- 10 (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:19 from nucleotide 114 to nucleotide 422;
- (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AJ27_1i deposited under accession number ATCC 98190;
- 15 (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AJ27_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AJ27_1i deposited under accession number ATCC 98190;
- 20 (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AJ27_1i deposited under accession number ATCC 98190;
- (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:20;
- (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:20 having biological activity;
- 25 (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and
- (k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

30 Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:19 from nucleotide 33 to nucleotide 422; the nucleotide sequence of SEQ ID NO:19 from nucleotide 114 to nucleotide 422; the nucleotide sequence of the full length protein coding sequence of clone AJ27_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AJ27_1i deposited under accession
35 number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full

length or mature protein encoded by the cDNA insert of clone AJ27_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group
5 consisting of:

- (a) the amino acid sequence of SEQ ID NO:20;
 - (b) fragments of the amino acid sequence of SEQ ID NO:20; and
 - (c) the amino acid sequence encoded by the cDNA insert of clone
AJ27_1i deposited under accession number ATCC 98190;
- 10 the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:20.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID
15 NO:22;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:22 from nucleotide 47 to nucleotide 517;
- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:22 from nucleotide 116 to nucleotide 517;
- 20 (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AJ142_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AJ142_1i deposited under accession number ATCC 98190;
- 25 (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AJ142_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AJ142_1i deposited under accession number ATCC 98190;
- 30 (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:23;
- (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:23 having biological activity;
- (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-
35 (g) above; and

(k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:22 from nucleotide 47 to nucleotide 517; the nucleotide sequence of SEQ ID NO:22 from nucleotide 116 to nucleotide 517; the nucleotide sequence of the full length protein coding sequence of clone AJ142_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AJ142_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AJ142_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:23;
- (b) fragments of the amino acid sequence of SEQ ID NO:23; and
- (c) the amino acid sequence encoded by the cDNA insert of clone

AJ142_1i deposited under accession number ATCC 98190; the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:23.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:24;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:24 from nucleotide 312 to nucleotide 417;
- (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AK604_1i deposited under accession number ATCC 98190;
- (d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AK604_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AK604_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AK604_1i deposited under accession number ATCC 98190;

(g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:25;

(h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:25 having biological activity;

5 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and

(j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:24
10 from nucleotide 312 to nucleotide 417; the nucleotide sequence of the full length protein coding sequence of clone AK604_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AK604_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AK604_1i
15 deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 20 (a) the amino acid sequence of SEQ ID NO:25;
(b) fragments of the amino acid sequence of SEQ ID NO:25; and
(c) the amino acid sequence encoded by the cDNA insert of clone AK604_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:25.

25 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:27;
(b) a polynucleotide comprising the nucleotide sequence of SEQ ID
30 NO:27 from nucleotide 76 to nucleotide 372;
(c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AK620_1i deposited under accession number ATCC 98190;
(d) a polynucleotide encoding the full length protein encoded by the
35 cDNA insert of clone AK620_1i deposited under accession number ATCC 98190;

(e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AK620_1i deposited under accession number ATCC 98190;

5 (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AK620_1i deposited under accession number ATCC 98190;

(g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:28;

(h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:28 having biological activity;

10 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and

(j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:27
15 from nucleotide 76 to nucleotide 372; the nucleotide sequence of the full length protein coding sequence of clone AK620_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AK620_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AK620_1i
20 deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 25 (a) the amino acid sequence of SEQ ID NO:28;
(b) fragments of the amino acid sequence of SEQ ID NO:28; and
(c) the amino acid sequence encoded by the cDNA insert of clone
AK620_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:28.

30 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID
NO:29;
(b) a polynucleotide comprising the nucleotide sequence of SEQ ID
35 NO:29 from nucleotide 367 to nucleotide 552;

- (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AK650_1i deposited under accession number ATCC 98190;
- 5 (d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AK650_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AK650_1i deposited under accession number ATCC 98190;
- 10 (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AK650_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:30;
- (h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:30 having biological activity;
- 15 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and
- (j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:29
20 from nucleotide 367 to nucleotide 552; the nucleotide sequence of the full length protein coding sequence of clone AK650_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AK650_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AK650_1i
25 deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 30 (a) the amino acid sequence of SEQ ID NO:30;
- (b) fragments of the amino acid sequence of SEQ ID NO:30; and
- (c) the amino acid sequence encoded by the cDNA insert of clone AK650_1i deposited under accession number ATCC 98190;
- the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:30.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:32;
- 5 (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:32 from nucleotide 116 to nucleotide 310;
- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:32 from nucleotide 173 to nucleotide 310;
- 10 (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AM226_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AM226_1i deposited under accession number ATCC 98190;
- 15 (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AM226_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AM226_1i deposited under accession number ATCC 98190;
- 20 (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:33;
- (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:33 having biological activity;
- (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and
- 25 (k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:32 from nucleotide 116 to nucleotide 310; the nucleotide sequence of SEQ ID NO:32 from nucleotide 173 to nucleotide 310; the nucleotide sequence of the full length protein coding sequence of clone AM226_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AM226_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AM226_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 5 (a) the amino acid sequence of SEQ ID NO:33;
- (b) fragments of the amino acid sequence of SEQ ID NO:33; and
- (c) the amino acid sequence encoded by the cDNA insert of clone AM226_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:33.

10 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:35;
- 15 (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:35 from nucleotide 281 to nucleotide 418;
- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:35 from nucleotide 353 to nucleotide 418;
- (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AR417_1i deposited under accession number ATCC 20 98190;
- (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AR417_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AR417_1i deposited under accession number ATCC 25 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AR417_1i deposited under accession number ATCC 98190;
- (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:36;
- 30 (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:36 having biological activity;
- (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and
- (k) a polynucleotide which encodes a species homologue of the protein 35 of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:35 from nucleotide 281 to nucleotide 418; the nucleotide sequence of SEQ ID NO:35 from nucleotide 353 to nucleotide 418; the nucleotide sequence of the full length protein coding sequence of clone AR417_1i deposited under accession number ATCC 98190; or the
5 nucleotide sequence of the mature protein coding sequence of clone AR417_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AR417_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a
10 protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:36;
 - (b) fragments of the amino acid sequence of SEQ ID NO:36; and
 - (c) the amino acid sequence encoded by the cDNA insert of clone
15 AR417_1i deposited under accession number ATCC 98190;
- the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:36.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- 20 (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:38;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:38 from nucleotide 496 to nucleotide 583;
- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID
25 NO:38 from nucleotide 565 to nucleotide 583;
- (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AU43_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide encoding the full length protein encoded by the
30 cDNA insert of clone AU43_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AU43_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA
35 insert of clone AU43_1i deposited under accession number ATCC 98190;

(h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:39;

(i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:39 having biological activity;

5 (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and

(k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:38
10 from nucleotide 496 to nucleotide 583; the nucleotide sequence of SEQ ID NO:38 from nucleotide 565 to nucleotide 583; the nucleotide sequence of the full length protein coding sequence of clone AU43_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AU43_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full
15 length or mature protein encoded by the cDNA insert of clone AU43_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 20 (a) the amino acid sequence of SEQ ID NO:39;
(b) fragments of the amino acid sequence of SEQ ID NO:39; and
(c) the amino acid sequence encoded by the cDNA insert of clone AU43_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein
25 comprises the amino acid sequence of SEQ ID NO:39.

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:41;
30 (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:41 from nucleotide 55 to nucleotide 405;
(c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:41 from nucleotide 148 to nucleotide 405;

- (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone AW60_1i deposited under accession number ATCC 98190;
- 5 (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone AW60_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone AW60_1i deposited under accession number ATCC 98190;
- 10 (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone AW60_1i deposited under accession number ATCC 98190;
- (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:42;
- (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:42 having biological activity;
- 15 (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and
- (k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:41 from nucleotide 55 to nucleotide 405; the nucleotide sequence of SEQ ID NO:41 from nucleotide 148 to nucleotide 405; the nucleotide sequence of the full length protein coding sequence of clone AW60_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone AW60_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone AW60_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 30 (a) the amino acid sequence of SEQ ID NO:42;
- (b) fragments of the amino acid sequence of SEQ ID NO:42; and
- (c) the amino acid sequence encoded by the cDNA insert of clone AW60_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:42.

35

In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:44;
- 5 (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:44 from nucleotide 337 to nucleotide 525;
- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:44 from nucleotide 406 to nucleotide 525;
- 10 (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone BA176_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BA176_1i deposited under accession number ATCC 98190;
- 15 (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BA176_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BA176_1i deposited under accession number ATCC 98190;
- 20 (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:45;
- (i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:45 having biological activity;
- (j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and
- 25 (k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:44 from nucleotide 337 to nucleotide 525; the nucleotide sequence of SEQ ID NO:44 from nucleotide 406 to nucleotide 525; the nucleotide sequence of the full length protein coding sequence of clone BA176_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone BA176_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone BA176_1i deposited under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:45;
- 5 (b) fragments of the amino acid sequence of SEQ ID NO:45; and
- (c) the amino acid sequence encoded by the cDNA insert of clone BA176_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:45.

10 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:47;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:47 from nucleotide 536 to nucleotide 628;
- 15 (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone BD140_1i deposited under accession number ATCC 98190;
- (d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BD140_1i deposited under accession number ATCC 98190;
- 20 (e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BD140_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BD140_1i deposited under accession number ATCC 98190;
- 25 (g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:48;
- (h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:48 having biological activity;
- 30 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and
- (j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:47 from nucleotide 536 to nucleotide 628; the nucleotide sequence of the full length protein

35

coding sequence of clone BD140_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone BD140_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone BD140_1i deposited
5 under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence of SEQ ID NO:48;
- 10 (b) fragments of the amino acid sequence of SEQ ID NO:48; and
- (c) the amino acid sequence encoded by the cDNA insert of clone BD140_1i deposited under accession number ATCC 98190;

the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:48.

15 In one embodiment, the present invention provides a composition comprising an isolated protein encoded by a polynucleotide selected from the group consisting of:

- (a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:50;
- (b) a polynucleotide comprising the nucleotide sequence of SEQ ID
20 NO:50 from nucleotide 303 to nucleotide 617;
- (c) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:50 from nucleotide 345 to nucleotide 617;
- (d) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone BD407_1i deposited under accession number ATCC
25 98190;
- (e) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BD407_1i deposited under accession number ATCC 98190;
- (f) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BD407_1i deposited under accession number ATCC
30 98190;
- (g) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BD407_1i deposited under accession number ATCC 98190;
- (h) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:51;

WO 98/14470

(i) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:51 having biological activity;

(j) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(g) above; and

5 (k) a polynucleotide which encodes a species homologue of the protein of (h) or (i) above.

Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:50 from nucleotide 303 to nucleotide 617; the nucleotide sequence of SEQ ID NO:50 from nucleotide 345 to nucleotide 617; the nucleotide sequence of the full length protein coding
10 sequence of clone BD407_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone BD407_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone BD407_1i deposited under accession number ATCC 98190. In yet other preferred embodiments, such
15 polynucleotide encodes a protein comprising the amino acid sequence of SEQ ID NO:51 from amino acid 1 to amino acid 32.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

20 (a) the amino acid sequence of SEQ ID NO:51;
(b) the amino acid sequence of SEQ ID NO:51 from amino acid 1 to amino acid 32;

(c) fragments of the amino acid sequence of SEQ ID NO:51; and
(d) the amino acid sequence encoded by the cDNA insert of clone
25 BD407_1i deposited under accession number ATCC 98190;
the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:51 or the amino acid sequence of SEQ ID NO:51 from amino acid 1 to amino acid 32.

In one embodiment, the present invention provides a composition comprising an
30 isolated protein encoded by a polynucleotide selected from the group consisting of:

(a) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:52;

(b) a polynucleotide comprising the nucleotide sequence of SEQ ID NO:52 from nucleotide 178 to nucleotide 534;

- (c) a polynucleotide comprising the nucleotide sequence of the full length protein coding sequence of clone BF290_1i deposited under accession number ATCC 98190;
- 5 (d) a polynucleotide encoding the full length protein encoded by the cDNA insert of clone BF290_1i deposited under accession number ATCC 98190;
- (e) a polynucleotide comprising the nucleotide sequence of the mature protein coding sequence of clone BF290_1i deposited under accession number ATCC 98190;
- 10 (f) a polynucleotide encoding the mature protein encoded by the cDNA insert of clone BF290_1i deposited under accession number ATCC 98190;
- (g) a polynucleotide encoding a protein comprising the amino acid sequence of SEQ ID NO:53;
- (h) a polynucleotide encoding a protein comprising a fragment of the amino acid sequence of SEQ ID NO:53 having biological activity;
- 15 (i) a polynucleotide which is an allelic variant of a polynucleotide of (a)-(f) above; and
- (j) a polynucleotide which encodes a species homologue of the protein of (g) or (h) above.
- 20 Preferably, such polynucleotide comprises the nucleotide sequence of SEQ ID NO:52 from nucleotide 178 to nucleotide 534; the nucleotide sequence of the full length protein coding sequence of clone BF290_1i deposited under accession number ATCC 98190; or the nucleotide sequence of the mature protein coding sequence of clone BF290_1i deposited under accession number ATCC 98190. In other preferred embodiments, the polynucleotide encodes the full length or mature protein encoded by the cDNA insert of clone BF290_1i deposited
- 25 under accession number ATCC 98190.

In other embodiments, the present invention provides a composition comprising a protein, wherein said protein comprises an amino acid sequence selected from the group consisting of:

- 30 (a) the amino acid sequence of SEQ ID NO:53;
- (b) fragments of the amino acid sequence of SEQ ID NO:53; and
- (c) the amino acid sequence encoded by the cDNA insert of clone BF290_1i deposited under accession number ATCC 98190;
- the protein being substantially free from other mammalian proteins. Preferably such protein comprises the amino acid sequence of SEQ ID NO:53.

Protein compositions of the present invention may further comprise a pharmaceutically acceptable carrier. Compositions comprising an antibody which specifically reacts with such protein are also provided by the present invention.

- Methods are also provided for preventing, treating or ameliorating a medical condition
- 5 which comprises administering to a mammalian subject a therapeutically effective amount of a composition comprising a protein of the present invention and a pharmaceutically acceptable carrier.

BRIEF DESCRIPTION OF FIGURES

- 10 Fig. 1 is a schematic representation of the pED6 and pNotS vectors used for deposit of clones disclosed herein.

Fig. 2 is an autoradiograph evidencing the expression of the following clone(s) disclosed herein: AE610_1i.

- Fig. 3 is an autoradiograph evidencing the expression of the following clone(s)
- 15 disclosed herein: AH106_1i, AM226_1i.

Fig. 4 is an autoradiograph evidencing the expression of the following clone(s) disclosed herein: AH196_1i.

Fig. 5 is an autoradiograph evidencing the expression of the following clone(s) disclosed herein: AI6_1i.

- 20 Fig. 6 is an autoradiograph evidencing the expression of the following clone(s) disclosed herein: AR417_1i.

Fig. 7 is an autoradiograph evidencing the expression of the following clone(s) disclosed herein: AW60_1i.

- Fig. 8 is an autoradiograph evidencing the expression of the following clone(s)
- 25 disclosed herein: BD140_1i.

Fig. 9 is an autoradiograph evidencing the expression of the following clone(s) disclosed herein: BF290_1i.

DETAILED DESCRIPTION

30 ISOLATED PROTEINS

- Nucleotide and amino acid sequences are reported below for each clone and protein disclosed in the present application. In some instances the sequences are preliminary and may include some incorrect or ambiguous bases or amino acids. The actual nucleotide sequence of each clone can readily be determined by sequencing of the deposited clone in accordance
- 35 with known methods. The predicted amino acid sequence (both full length and mature) can

then be determined from such nucleotide sequence. The amino acid sequence of the protein encoded by a particular clone can also be determined by expression of the clone in a suitable host cell, collecting the protein and determining its sequence.

For each disclosed protein applicants have identified what they have determined to be the reading frame best identifiable with sequence information available at the time of filing. Because of the partial ambiguity in reported sequence information, reported protein sequences include "Xaa" designators. These "Xaa" designators indicate either (1) a residue which cannot be identified because of nucleotide sequence ambiguity or (2) a stop codon in the determined nucleotide sequence where applicants believe one should not exist (if the nucleotide sequence were determined more accurately).

As used herein a "secreted" protein is one which, when expressed in a suitable host cell, is transported across or through a membrane, including transport as a result of signal sequences in its amino acid sequence. "Secreted" proteins include without limitation proteins secreted wholly (e.g., soluble proteins) or partially (e.g., receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins which are transported across the membrane of the endoplasmic reticulum.

Protein "AE402_1i"

One protein of the present invention has been identified as protein "AE402_1i". A partial cDNA clone encoding AE402_1i was first isolated from a murine adult spleen cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yh02h12.r1 Homo sapiens cDNA clone 42238 5'" (R60758, BlastN) and "yh02h12.s1 Homo sapiens cDNA clone 42238 3'" (R60759, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AE402_1i".

Applicants' methods identified clone AE402_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AE402_1i as presently determined is reported in SEQ ID NO:1. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AE402_1i protein corresponding to the foregoing

nucleotide sequence is reported in SEQ ID NO:2. Additional nucleotide sequence from the 3' portion of AE402_1i, including the polyA tail, is reported in SEQ ID NO:3.

Protein "AE610_1i"

5 One protein of the present invention has been identified as protein "AE610_1i". A partial cDNA clone encoding AE610_1i was first isolated from a murine adult spleen cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least
10 some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yf19g02.r1 Homo sapiens cDNA" (R08399, Fasta), "yw68d09.s1 Homo sapiens cDNA clone 257393 3'" (N27174, BlastN), "yi10a04.r1 Homo sapiens cDNA" (R62698, Fasta) and "yh78e10.s1 Homo sapiens cDNA clone 135882 3'" (R33815, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor
15 of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AE610_1i".

Applicants' methods identified clone AE610_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AE610_1i as presently determined is
20 reported in SEQ ID NO:4. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AE610_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:5. Amino acids 1 to 87 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 88. Additional nucleotide sequence from the 3' portion of AE610_1i, including the polyA tail.
25 is reported in SEQ ID NO:6.

Protein "AH106_1i"

One protein of the present invention has been identified as protein "AH106_1i". A partial cDNA clone encoding AH106_1i was first isolated from a murine fetal thymus cDNA
30 library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with an EST reported by the I.M.A.G.E. Consortium identified at GenBank accession number T81127. The human cDNA clone corresponding to the EST database entry
35 was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E.

Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AH106_1i".

Applicants' methods identified clone AH106_1i as encoding a secreted protein.

5 The nucleotide sequence of AH106_1i as presently determined is reported in SEQ ID NO:7. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AH106_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:8..

10 Protein "AH196_1i"

One protein of the present invention has been identified as protein "AH196_1i". A partial cDNA clone encoding AH196_1i was first isolated from a murine fetal thymus cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank
15 database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yj12f04.r1 Homo sapiens cDNA clone 148543 5'" (H12523, BlastN) and "yj12f04.s1 Homo sapiens cDNA clone 148543 3'" (H12470, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the
20 I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AH196_1i".

Applicants' methods identified clone AH196_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AH196_1i as presently determined is
25 reported in SEQ ID NO:9. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AH196_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:10. Additional nucleotide sequence from the 3' portion of AH196_1i, including the polyA tail, is reported in SEQ ID NO:11.

30 Protein "AI6_1i"

One protein of the present invention has been identified as protein "AI6_1i". A partial cDNA clone encoding AI6_1i was first isolated from a human blood cell (Th1 or Th2) cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank
35 database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least

some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yj42h04.r1 Homo sapiens cDNA" (H03613, Fasta) and "yx60f10.s1 Homo sapiens cDNA clone 266155 3'" (N21637, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AI6_1i".

Applicants' methods identified clone AI6_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AI6_1i as presently determined is reported in SEQ ID NO:12. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AI6_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:13. Additional nucleotide sequence from the 3' portion of AI6_1i, including the polyA tail, is reported in SEQ ID NO:14.

15 Protein "AJ13_1i"

One protein of the present invention has been identified as protein "AJ13_1i". A partial cDNA clone encoding AJ13_1i was first isolated from a human adult testes cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yo61h02.r1 Homo sapiens cDNA clone 182451 5'" (H42116, BlastN), "yr84a08.r1 Homo sapiens cDNA clone 211958 5'" (H75363, BlastN) and "yg83h03.s1 Homo sapiens cDNA clone 40148 3'" (R53978, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AJ13_1i".

Applicants' methods identified clone AJ13_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AJ13_1i as presently determined is reported in SEQ ID NO:15. An additional internal nucleotide sequence from AJ13_1i as presently determined is reported in SEQ ID NO:16. What applicants believe is the proper reading frame and the predicted amino acid sequence encoded by such internal sequence is reported in SEQ ID NO:17. Additional nucleotide sequence from the 3' portion of AJ13_1i, including the polyA tail, is reported in SEQ ID NO:18.

Protein "AJ27_1i"

One protein of the present invention has been identified as protein "AJ27_1i". A partial cDNA clone encoding AJ27_1i was first isolated from a human adult testes cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yx25h01.r1 Homo sapiens cDNA clone 262897 5'" (N28373, BlastN) and "yx62d05.r1 Homo sapiens cDNA clone 266313 5'" (N35654, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AJ27_1i".

Applicants' methods identified clone AJ27_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AJ27_1i as presently determined is reported in SEQ ID NO:19. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AJ27_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:20. Amino acids 1 to 27 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 28. Additional nucleotide sequence from the 3' portion of AJ27_1i, including the polyA tail, is reported in SEQ ID NO:21.

Protein "AJ142_1i"

One protein of the present invention has been identified as protein "AJ142_1i". A partial cDNA clone encoding AJ142_1i was first isolated from a human adult testes cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yq85b12.r1 Homo sapiens cDNA clone 202559 5'" (H53268, BlastN) and "yq85b12.s1 Homo sapiens cDNA clone 202559 3'" (H53269, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AJ142_1i".

Applicants' methods identified clone AJ142_1i as encoding a secreted protein.

The nucleotide sequence of AJ142_1i as presently determined is reported in SEQ ID NO:22. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AJ142_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:23. Amino acids 1 to 23 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 24.

Protein "AK604_1i"

One protein of the present invention has been identified as protein "AK604_1i". A partial cDNA clone encoding AK604_1i was first isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with an EST reported by the I.M.A.G.E. Consortium identified as "yc80g11.r1 Homo sapiens cDNA clone 22157 5'" (T64857, BlastN). The sequence also showed at least some identity with a partial cDNA sequence identified as "H. sapiens partial cDNA sequence; clone c-1pg11" (Z40033, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AK604_1i".

Applicants' methods identified clone AK604_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AK604_1i as presently determined is reported in SEQ ID NO:24. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AK604_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:25. Additional nucleotide sequence from the 3' portion of AK604_1i, including the polyA tail, is reported in SEQ ID NO:26.

Protein "AK620_1i"

One protein of the present invention has been identified as protein "AK620_1i". A partial cDNA clone encoding AK620_1i was first isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "ye7607.r1

Homo sapiens cDNA clone 123684 5'" (R02637, BlastN) and "yx90e05.s1 Homo sapiens cDNA clone 269024 3'" (N26101, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and
5 determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AK620_1i".

Applicants' methods identified clone AK620_1i as encoding a secreted protein.

The nucleotide sequence of AK620_1i as presently determined is reported in SEQ ID NO:27. What applicants believe is the proper reading frame and the predicted amino acid
10 sequence of the AK620_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:28..

Protein "AK650_1i"

One protein of the present invention has been identified as protein "AK650_1i". A
15 partial cDNA clone encoding AK650_1i was first isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yp60g06.r1
20 Homo sapiens cDNA clone 191866 5'" (H40407, BlastN) and "yp60g06.s1 Homo sapiens cDNA clone 191866 3'" (H40350, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail).
25 This full-length clone is also referred to herein as "AK650_1i".

Applicants' methods identified clone AK650_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AK650_1i as presently determined is reported in SEQ ID NO:29. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AK650_1i protein corresponding to the foregoing
30 nucleotide sequence is reported in SEQ ID NO:30. Additional nucleotide sequence from the 3' portion of AK650_1i, including the polyA tail, is reported in SEQ ID NO:31.

Protein "AM226_1i"

One protein of the present invention has been identified as protein "AM226_1i". A
35 partial cDNA clone encoding AM226_1i was first isolated from a human fetal kidney cDNA

library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yf09a01.r1
5 Homo sapiens cDNA clone 126312 5'" (R06469, BlastN) and "yy49b06.s1 Homo sapiens cDNA clone 276851 3'" (N39415, BlastN). The sequence also showed some similarity with bovine osteoinductive factor (OIF) (M37974, BlastN), with which it may share some activity. The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone
10 received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AM226_1i".

Applicants' methods identified clone AM226_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AM226_1i as presently determined is
15 reported in SEQ ID NO:32. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AM226_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:33. Amino acids 1 to 19 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid
20. Additional nucleotide sequence from the 3' portion of AM226_1i, including the polyA tail,
20 is reported in SEQ ID NO:34.

Protein "AR417_1i"

One protein of the present invention has been identified as protein "AR417_1i". A partial cDNA clone encoding AR417_1i was first isolated from a human adult retina cDNA
25 library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified at GenBank accession numbers R18973, R42209 ("yf89g09.s1 Homo sapiens cDNA clone 29781 3'"),
30 R12416 ("yf56a02.r1 Homo sapiens cDNA clone 26106 5'") and R15309 ("yf89g09.r1 Homo sapiens cDNA"). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is
35 also referred to herein as "AR417_1i".

Applicants' methods identified clone AR417_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AR417_1i as presently determined is reported in SEQ ID NO:35. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AR417_1i protein corresponding to the foregoing
5 nucleotide sequence is reported in SEQ ID NO:36. Amino acids 1 to 24 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 25. Additional nucleotide sequence from the 3' portion of AR417_1i, including the polyA tail, is reported in SEQ ID NO:37.

10 Protein "AU43_1i"

One protein of the present invention has been identified as protein "AU43_1i". A partial cDNA clone encoding AU43_1i was first isolated from a human adult testes cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank
15 database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "y149f07.r1 Homo sapiens cDNA clone 142597 5'" (R70850, BlastN) and "yd68e02.s1 Homo sapiens cDNA clone 113402 3'" (T78464, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the
20 I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AU43_1i".

Applicants' methods identified clone AU43_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AU43_1i as presently determined is
25 reported in SEQ ID NO:38. What applicants believe is the proper reading frame and the predicted amino acid sequence of the AU43_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:39. Amino acids 1 to 23 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 24. Additional nucleotide sequence from the 3' portion of AU43_1i, including the polyA tail,
30 is reported in SEQ ID NO:40.

Protein "AW60_1i"

One protein of the present invention has been identified as protein "AW60_1i". A partial cDNA clone encoding AW60_1i was first isolated from a human ovary (PA-1
35 teratocarcinoma) cDNA library using methods which are selective for cDNAs encoding

secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "ym57f11.r1 Homo sapiens cDNA clone 52343 5'" (H23492, BlastN),
5 "ym57f08.r1 Homo sapiens cDNA" (H23390, Fasta) and "ym57f11.s1 Homo sapiens cDNA clone 52343 3'" (H23494, BlastN). The sequence also showed at least some identity with a sequence identified as "Homo sapiens clone S31i125" (L40397, Fasta). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the
10 distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "AW60_1i".

Applicants' methods identified clone AW60_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of AW60_1i as presently determined is reported in SEQ ID NO:41. What applicants believe is the proper reading frame and the
15 predicted amino acid sequence of the AW60_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:42. Amino acids 1 to 31 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 32. Additional nucleotide sequence from the 3' portion of AW60_1i, including the polyA tail, is reported in SEQ ID NO:43.

20

Protein "BA176_1i"

One protein of the present invention has been identified as protein "BA176_1i". A partial cDNA clone encoding BA176_1i was first isolated from a human adult placenta cDNA library using methods which are selective for cDNAs encoding secreted proteins. The
25 nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yi75g11.r1 Homo sapiens cDNA" (R77409, Fasta), "yj50b12.r1 Homo sapiens cDNA" (H03089, Fasta) and "yi75g11.s1 Homo sapiens cDNA clone 145124 3'" (R77410, BlastN). The human cDNA
30 clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "BA176_1i".

Applicants' methods identified clone BA176_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of BA176_1i as presently determined is reported in SEQ ID NO:44. What applicants believe is the proper reading frame and the predicted amino acid sequence of the BA176_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:45. Amino acids 1 to 23 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 24. Additional nucleotide sequence from the 3' portion of BA176_1i, including the polyA tail, is reported in SEQ ID NO:46.

Protein "BD140_1i"

One protein of the present invention has been identified as protein "BD140_1i". A partial cDNA clone encoding BD140_1i was first isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yn98c02.r1 Homo sapiens cDNA" (H43507, Fasta), "yn67g04.r1 Homo sapiens cDNA" (H22693, Fasta) and "yn82e07.s1 Homo sapiens cDNA clone 174948 3'" (H38408, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "BD140_1i".

Applicants' methods identified clone BD140_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of BD140_1i as presently determined is reported in SEQ ID NO:47. What applicants believe is the proper reading frame and the predicted amino acid sequence of the BD140_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:48. Additional nucleotide sequence from the 3' portion of BD140_1i, including the polyA tail, is reported in SEQ ID NO:49.

Protein "BD407_1i"

One protein of the present invention has been identified as protein "BD407_1i". A partial cDNA clone encoding BD407_1i was first isolated from a human fetal kidney cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "ys65a05.r1

Homo sapiens cDNA" (H84524, Fasta) and "yz15h02.s1 Homo sapiens cDNA clone 283155 3'" (N51349, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full
5 length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "BD407_1i".

Applicants' methods identified clone BD407_1i as encoding a secreted protein.

The nucleotide sequence of BD407_1i as presently determined is reported in SEQ ID NO:50. What applicants believe is the proper reading frame and the predicted amino acid
10 sequence of the BD407_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:51. Amino acids 1 to 14 are the predicted leader/signal sequence, with the predicted mature amino acid sequence beginning at amino acid 15.

Protein "BF290_1i"

15 One protein of the present invention has been identified as protein "BF290_1i". A partial cDNA clone encoding BF290_1i was first isolated from a human fetal brain cDNA library using methods which are selective for cDNAs encoding secreted proteins. The nucleotide sequence of such partial cDNA was determined and searched against the GenBank database using BLASTA/BLASTX and FASTA search protocols. The search revealed at least
20 some identity with ESTs reported by the I.M.A.G.E. Consortium identified as "yh10f04.r1 Homo sapiens cDNA" (R61165, Fasta) and "yy35d12.s1 Homo sapiens cDNA clone 273239 3'" (N33175, BlastN). The human cDNA clone corresponding to the EST database entry was ordered from Genome Systems, Inc., St. Louis, Mo, a distributor of the I.M.A.G.E. Consortium library. The clone received from the distributor was examined and determined to be a full
25 length clone, including a 5' end and 3' UTR (including a polyA tail). This full-length clone is also referred to herein as "BF290_1i".

Applicants' methods identified clone BF290_1i as encoding a secreted protein.

The nucleotide sequence of the 5' portion of BF290_1i as presently determined is reported in SEQ ID NO:52. What applicants believe is the proper reading frame and the
30 predicted amino acid sequence of the BF290_1i protein corresponding to the foregoing nucleotide sequence is reported in SEQ ID NO:53. Additional nucleotide sequence from the 3' portion of BF290_1i, including the polyA tail, is reported in SEQ ID NO:54.

Deposit of Clones

Clones AE402_1i, AE610_1i, AH106_1i, AH196_1i, AI6_1i, AJ13_1i, AJ27_1i, AJ142_1i, AK604_1i, AK620_1i, AK650_1i, AM226_1i, AR417_1i, AU43_1i, AW60_1i, BA176_1i, BD140_1i, BD407_1i and BF290_1i were deposited on October 2, 1996 with the American Type Culture Collection under accession number ATCC 98190, from which each clone comprising a particular polynucleotide is obtainable. Each clone has been transfected into separate bacterial cells (*E. coli*) in this composite deposit.

Each clone can be removed from the vector in which it was deposited by performing an EcoRI/NotI digestion (5' cite, EcoRI; 3' cite, NotI) to produce the appropriate fragment for such clone. Each clone was deposited in either the pED6 or pNotS vector depicted in Fig. 1. In some instances, the deposited clone can become "flipped" (i.e., in the reverse orientation) in the deposited isolate. In such instances, the cDNA insert can still be isolated by digestion with EcoRI and NotI. However, NotI will then produce the 5' cite and EcoRI will produce the 3' cite for placement of the cDNA in proper orientation for expression in a suitable vector. The cDNA may also be expressed from the vectors in which they were deposited.

Bacterial cells containing a particular clone can be obtained from the composite deposit as follows:

An oligonucleotide probe or probes should be designed to the sequence that is known for that particular clone. This sequence can be derived from the sequences provided herein, or from a combination of those sequences.

In the sequences listed above which include an N at position 2, that position is occupied in preferred probes/primers by a biotinylated phosphoramidite residue rather than a nucleotide (such as , for example, that produced by use of biotin phosphoramidite (1-dimethoxytrityloxy-2-(N-biotinyl-4-aminobutyl)-propyl-3-O-(2-cyanoethyl)-(N,N-diisopropyl)-phosphoramidite) (Glen Research, cat. no. 10-1953)).

The design of the oligonucleotide probe should preferably follow these parameters:

- (a) It should be designed to an area of the sequence which has the fewest ambiguous bases ("N's"), if any;
- (b) It should be designed to have a T_m of approx. 80 ° C (assuming 2° for each A or T and 4 degrees for each G or C).

The oligonucleotide should preferably be labeled with $g\text{-}^{32}\text{P}$ ATP (specific activity 6000 Ci/mmole) and T4 polynucleotide kinase using commonly employed techniques for labeling oligonucleotides. Other labeling techniques can also be used. Unincorporated label should preferably be removed by gel filtration chromatography or other established methods. The amount of radioactivity incorporated into the probe should be quantitated by measurement in

a scintillation counter. Preferably, specific activity of the resulting probe should be approximately 4×10^6 dpm/pmole.

The bacterial culture containing the pool of full-length clones should preferably be thawed and 100 μ l of the stock used to inoculate a sterile culture flask containing 25 ml of sterile L-broth containing ampicillin at 100 μ g/ml. The culture should preferably be grown to saturation at 37°C, and the saturated culture should preferably be diluted in fresh L-broth. Aliquots of these dilutions should preferably be plated to determine the dilution and volume which will yield approximately 5000 distinct and well-separated colonies on solid bacteriological media containing L-broth containing ampicillin at 100 μ g/ml and agar at 1.5% in a 150 mm petri dish when grown overnight at 37°C. Other known methods of obtaining distinct, well-separated colonies can also be employed.

Standard colony hybridization procedures should then be used to transfer the colonies to nitrocellulose filters and lyse, denature and bake them.

The filter is then preferably incubated at 65°C for 1 hour with gentle agitation in 6X SSC (20X stock is 175.3 g NaCl/liter, 88.2 g Na citrate/liter, adjusted to pH 7.0 with NaOH) containing 0.5% SDS, 100 μ g/ml of yeast RNA, and 10 mM EDTA (approximately 10 mL per 150 mm filter). Preferably, the probe is then added to the hybridization mix at a concentration greater than or equal to 1×10^6 dpm/mL. The filter is then preferably incubated at 65°C with gentle agitation overnight. The filter is then preferably washed in 500 mL of 2X SSC/0.5% SDS at room temperature without agitation, preferably followed by 500 mL of 2X SSC/0.1% SDS at room temperature with gentle shaking for 15 minutes. A third wash with 0.1X SSC/0.5% SDS at 65°C for 30 minutes to 1 hour is optional. The filter is then preferably dried and subjected to autoradiography for sufficient time to visualize the positives on the X-ray film. Other known hybridization methods can also be employed.

The positive colonies are picked, grown in culture, and plasmid DNA isolated using standard procedures. The clones can then be verified by restriction analysis, hybridization analysis, or DNA sequencing.

Fragments of the proteins of the present invention which are capable of exhibiting biological activity are also encompassed by the present invention. Fragments of the protein may be in linear form or they may be cyclized using known methods, for example, as described in H.U. Saragovi, *et al.*, Bio/Technology 10, 773-778 (1992) and in R.S. McDowell, *et al.*, J. Amer. Chem. Soc. 114, 9245-9253 (1992), both of which are incorporated herein by reference. Such fragments may be fused to carrier molecules such as immunoglobulins for many purposes, including increasing the valency of protein binding sites. For example, fragments of the protein may be fused through "linker" sequences to the Fc portion of an

immunoglobulin. For a bivalent form of the protein, such a fusion could be to the Fc portion of an IgG molecule. Other immunoglobulin isotypes may also be used to generate such fusions. For example, a protein - IgM fusion would generate a decavalent form of the protein of the invention.

5 The present invention also provides both full-length and mature forms of the disclosed proteins. The full-length form of the such proteins is identified in the sequence listing by translation of the nucleotide sequence of each disclosed clone. The mature form of such protein may be obtained by expression of the disclosed full-length polynucleotide (preferably those deposited with ATCC) in a suitable mammalian cell or other host cell. The sequence of
10 the mature form of the protein may also be determinable from the amino acid sequence of the full-length form.

 Where the protein of the present invention is membrane-bound (e.g., is a receptor), the present invention also provides for soluble forms of such protein. In such forms part or all of the intracellular and transmembrane domains of the protein are deleted such that the protein
15 is fully secreted from the cell in which it is expressed. The intracellular and transmembrane domains of proteins of the invention can be identified in accordance with known techniques for determination of such domains from sequence information.

 Species homologs of the disclosed proteins are also provided by the present invention. Species homologs may be isolated and identified by making suitable probes or primers from
20 the sequences provided herein and screening a suitable nucleic acid source from the desired species.

 The invention also encompasses allelic variants of the disclosed proteins; that is, naturally-occurring alternative forms of the isolated proteins which are identical, homologous or related to that encoded by the polynucleotides disclosed herein.

25 The isolated polynucleotide encoding the protein of the invention may be operably linked to an expression control sequence such as the pMT2 or pED expression vectors disclosed in Kaufman *et al.*, Nucleic Acids Res. 19, 4485-4490 (1991), in order to produce the protein recombinantly. Many suitable expression control sequences are known in the art. General methods of expressing recombinant proteins are also known and are exemplified in
30 R. Kaufman, Methods in Enzymology 185, 537-566 (1990). As defined herein "operably linked" means that the isolated polynucleotide of the invention and an expression control sequence are situated within a vector or cell in such a way that the protein is expressed by a host cell which has been transformed (transfected) with the ligated polynucleotide/expression control sequence.

A number of types of cells may act as suitable host cells for expression of the protein. Mammalian host cells include, for example, monkey COS cells, Chinese Hamster Ovary (CHO) cells, human kidney 293 cells, human epidermal A431 cells, human Colo205 cells, 3T3 cells, CV-1 cells, other transformed primate cell lines, normal diploid cells, cell strains derived
5 from in vitro culture of primary tissue, primary explants, HeLa cells, mouse L cells, BHK, HL-60, U937, HaK or Jurkat cells.

Alternatively, it may be possible to produce the protein in lower eukaryotes such as yeast or in prokaryotes such as bacteria. Potentially suitable yeast strains include *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Kluyveromyces* strains, *Candida*, or
10 any yeast strain capable of expressing heterologous proteins. Potentially suitable bacterial strains include *Escherichia coli*, *Bacillus subtilis*, *Salmonella typhimurium*, or any bacterial strain capable of expressing heterologous proteins. If the protein is made in yeast or bacteria, it may be necessary to modify the protein produced therein, for example by phosphorylation or glycosylation of the appropriate sites, in order to obtain the functional protein. Such
15 covalent attachments may be accomplished using known chemical or enzymatic methods.

The protein may also be produced by operably linking the isolated polynucleotide of the invention to suitable control sequences in one or more insect expression vectors, and employing an insect expression system. Materials and methods for baculovirus/insect cell expression systems are commercially available in kit form from, *e.g.*, Invitrogen, San Diego,
20 California, U.S.A. (the MaxBac® kit), and such methods are well known in the art, as described in Summers and Smith, Texas Agricultural Experiment Station Bulletin No. 1555 (1987), incorporated herein by reference. As used herein, an insect cell capable of expressing a polynucleotide of the present invention is "transformed."

The protein of the invention may be prepared by culturing transformed host cells under
25 culture conditions suitable to express the recombinant protein. The resulting expressed protein may then be purified from such culture (*i.e.*, from culture medium or cell extracts) using known purification processes, such as gel filtration and ion exchange chromatography. The purification of the protein may also include an affinity column containing agents which will bind to the protein; one or more column steps over such affinity resins as concanavalin A-
30 agarose, heparin-toyopearl® or Cibacrom blue 3GA Sepharose®; one or more steps involving hydrophobic interaction chromatography using such resins as phenyl ether, butyl ether, or propyl ether; or immunoaffinity chromatography.

Alternatively, the protein of the invention may also be expressed in a form which will facilitate purification. For example, it may be expressed as a fusion protein, such as those of
35 maltose binding protein (MBP), glutathione-S-transferase (GST) or thioredoxin (TRX). Kits

for expression and purification of such fusion proteins are commercially available from New England BioLab (Beverly, MA), Pharmacia (Piscataway, NJ) and InVitrogen, respectively. The protein can also be tagged with an epitope and subsequently purified by using a specific antibody directed to such epitope. One such epitope ("Flag") is commercially available from
5 Kodak (New Haven, CT).

Finally, one or more reverse-phase high performance liquid chromatography (RP-HPLC) steps employing hydrophobic RP-HPLC media, e.g., silica gel having pendant methyl or other aliphatic groups, can be employed to further purify the protein. Some or all of the foregoing purification steps, in various combinations, can also be employed to provide a
10 substantially homogeneous isolated recombinant protein. The protein thus purified is substantially free of other mammalian proteins and is defined in accordance with the present invention as an "isolated protein."

The protein of the invention may also be expressed as a product of transgenic animals, e.g., as a component of the milk of transgenic cows, goats, pigs, or sheep which are
15 characterized by somatic or germ cells containing a nucleotide sequence encoding the protein.

The protein may also be produced by known conventional chemical synthesis. Methods for constructing the proteins of the present invention by synthetic means are known to those skilled in the art. The synthetically-constructed protein sequences, by virtue of sharing primary, secondary or tertiary structural and/or conformational characteristics with proteins
20 may possess biological properties in common therewith, including protein activity. Thus, they may be employed as biologically active or immunological substitutes for natural, purified proteins in screening of therapeutic compounds and in immunological processes for the development of antibodies.

The proteins provided herein also include proteins characterized by amino acid
25 sequences similar to those of purified proteins but into which modification are naturally provided or deliberately engineered. For example, modifications in the peptide or DNA sequences can be made by those skilled in the art using known techniques. Modifications of interest in the protein sequences may include the alteration, substitution, replacement, insertion or deletion of a selected amino acid residue in the coding sequence. For example, one or more
30 of the cysteine residues may be deleted or replaced with another amino acid to alter the conformation of the molecule. Techniques for such alteration, substitution, replacement, insertion or deletion are well known to those skilled in the art (see, e.g., U.S. Patent No. 4,518,584). Preferably, such alteration, substitution, replacement, insertion or deletion retains the desired activity of the protein.

Other fragments and derivatives of the sequences of proteins which would be expected to retain protein activity in whole or in part and may thus be useful for screening or other immunological methodologies may also be easily made by those skilled in the art given the disclosures herein. Such modifications are believed to be encompassed by the present invention.

USES AND BIOLOGICAL ACTIVITY

The proteins of the present invention are expected to exhibit one or more of the uses or biological activities (including those associated with assays cited herein) identified below.

Uses or activities described for proteins of the present invention may be provided by administration or use of such proteins or by administration or use of polynucleotides encoding such proteins (such as, for example, in gene therapies or vectors suitable for introduction of DNA).

Research Uses and Utilities

The proteins provided by the present invention can similarly be used in assay to determine biological activity, including in a panel of multiple proteins for high-throughput screening; to raise antibodies or to elicit another immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the protein (or its receptor) in biological fluids; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state); and, of course, to isolate correlative receptors or ligands. Where the protein binds or potentially binds to another protein (such as, for example, in a receptor-ligand interaction), the protein can be used to identify the other protein with which binding occurs or to identify inhibitors of the binding interaction. Proteins involved in these binding interactions can also be used to screen for peptide or small molecule inhibitors or agonists of the binding interaction.

Any or all of these research utilities are capable of being developed into reagent grade or kit format for commercialization as research products.

Methods for performing the uses listed above are well known to those skilled in the art. References disclosing such methods include without limitation "Molecular Cloning: A Laboratory Manual", 2d ed., Cold Spring Harbor Laboratory Press, Sambrook, J., E.F. Fritsch and T. Maniatis eds., 1989, and "Methods in Enzymology: Guide to Molecular Cloning Techniques", Academic Press, Berger, S.L. and A.R. Kimmel eds., 1987.

Nutritional Uses

Proteins of the present invention can also be used as nutritional sources or supplements. Such uses include without limitation use as a protein or amino acid supplement, use as a carbon source, use as a nitrogen source and use as a source of carbohydrate. In such cases the protein of the invention can be added to the feed of a particular organism or can be administered as a separate solid or liquid preparation, such as in the form of powder, pills, solutions, suspensions or capsules. In the case of microorganisms, the protein of the invention can be added to the medium in or on which the microorganism is cultured.

Cytokine and Cell Proliferation/Differentiation Activity

A protein of the present invention may exhibit cytokine, cell proliferation (either inducing or inhibiting) or cell differentiation (either inducing or inhibiting) activity or may induce production of other cytokines in certain cell populations. Many protein factors discovered to date, including all known cytokines, have exhibited activity in one or more factor dependent cell proliferation assays, and hence the assays serve as a convenient confirmation of cytokine activity. The activity of a protein of the present invention is evidenced by any one of a number of routine factor dependent cell proliferation assays for cell lines including, without limitation, 32D, DA2, DA1G, T10, B9, B9/11, BaF3, MC9/G, M+ (preB M+), 2E8, RB5, DA1, 123, T1165, HT2, CTLL2, TF-1, Mo7e and CMK.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for T-cell or thymocyte proliferation include without limitation those described in: *Current Protocols in Immunology*, Ed by J. E. Coligan, A.M. Kruisbeek, D.H. Margulies, E.M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, In Vitro assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, Immunologic studies in Humans); Takai et al., *J. Immunol.* 137:3494-3500, 1986; Bertagnolli et al., *J. Immunol.* 145:1706-1712, 1990; Bertagnolli et al., *Cellular Immunology* 133:327-341, 1991; Bertagnolli, et al., *J. Immunol.* 149:3778-3783, 1992; Bowman et al., *J. Immunol.* 152: 1756-1761, 1994.

Assays for cytokine production and/or proliferation of spleen cells, lymph node cells or thymocytes include, without limitation, those described in: Polyclonal T cell stimulation, Kruisbeek, A.M. and Shevach, E.M. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 3.12.1-3.12.14, John Wiley and Sons, Toronto. 1994; and Measurement of mouse and human Interferon γ , Schreiber, R.D. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 6.8.1-6.8.8, John Wiley and Sons, Toronto. 1994.

Assays for proliferation and differentiation of hematopoietic and lymphopoietic cells include, without limitation, those described in: Measurement of Human and Murine Interleukin 2 and Interleukin 4, Bottomly, K., Davis, L.S. and Lipsky, P.E. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 6.3.1-6.3.12, John Wiley and Sons, Toronto. 1991; deVries et al., J. Exp. Med. 173:1205-1211, 1991; Moreau et al., Nature 336:690-692, 1988; Greenberger et al., Proc. Natl. Acad. Sci. U.S.A. 80:2931-2938, 1983; Measurement of mouse and human interleukin 6 - Nordan, R. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 6.6.1-6.6.5, John Wiley and Sons, Toronto. 1991; Smith et al., Proc. Natl. Acad. Sci. U.S.A. 83:1857-1861, 1986; Measurement of human Interleukin 11 - Bennett, F., Giannotti, J., Clark, S.C. and Turner, K. J. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 6.15.1 John Wiley and Sons, Toronto. 1991; Measurement of mouse and human Interleukin 9 - Ciarletta, A., Giannotti, J., Clark, S.C. and Turner, K.J. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 6.13.1, John Wiley and Sons, Toronto. 1991.

Assays for T-cell clone responses to antigens (which will identify, among others, proteins that affect APC-T cell interactions as well as direct T-cell effects by measuring proliferation and cytokine production) include, without limitation, those described in: Current Protocols in Immunology, Ed by J. E. Coligan, A.M. Kruisbeek, D.H. Margulies, E.M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3. In Vitro assays for Mouse Lymphocyte Function; Chapter 6, Cytokines and their cellular receptors; Chapter 7, Immunologic studies in Humans); Weinberger et al., Proc. Natl. Acad. Sci. USA 77:6091-6095, 1980; Weinberger et al., Eur. J. Immun. 11:405-411, 1981; Takai et al., J. Immunol. 137:3494-3500, 1986; Takai et al., J. Immunol. 140:508-512, 1988.

Immune Stimulating or Suppressing Activity

A protein of the present invention may also exhibit immune stimulating or immune suppressing activity, including without limitation the activities for which assays are described herein. A protein may be useful in the treatment of various immune deficiencies and disorders (including severe combined immunodeficiency (SCID)), e.g., in regulating (up or down) growth and proliferation of T and/or B lymphocytes, as well as effecting the cytolytic activity of NK cells and other cell populations. These immune deficiencies may be genetic or be caused by viral (e.g., HIV) as well as bacterial or fungal infections, or may result from autoimmune disorders. More specifically, infectious diseases caused by viral, bacterial, fungal or other infection may be treatable using a protein of the present invention, including infections by HIV, hepatitis viruses, herpesviruses, mycobacteria, *Leishmania* spp., malaria

spp. and various fungal infections such as candidiasis. Of course, in this regard, a protein of the present invention may also be useful where a boost to the immune system generally may be desirable, *i.e.*, in the treatment of cancer.

Autoimmune disorders which may be treated using a protein of the present invention
5 include, for example, connective tissue disease, multiple sclerosis, systemic lupus erythematosus, rheumatoid arthritis, autoimmune pulmonary inflammation, Guillain-Barre syndrome, autoimmune thyroiditis, insulin dependent diabetes mellitus, myasthenia gravis, graft-versus-host disease and autoimmune inflammatory eye disease. Such a protein of the present invention may also be useful in the treatment of allergic reactions and conditions,
10 such as asthma (particularly allergic asthma) or other respiratory problems. Other conditions, in which immune suppression is desired (including, for example, organ transplantation), may also be treatable using a protein of the present invention.

Using the proteins of the invention it may also be possible to immune responses, in a number of ways. Down regulation may be in the form of inhibiting or blocking an immune
15 response already in progress or may involve preventing the induction of an immune response. The functions of activated T cells may be inhibited by suppressing T cell responses or by inducing specific tolerance in T cells, or both. Immunosuppression of T cell responses is generally an active, non-antigen-specific, process which requires continuous exposure of the T cells to the suppressive agent. Tolerance, which involves inducing non-responsiveness or
20 anergy in T cells, is distinguishable from immunosuppression in that it is generally antigen-specific and persists after exposure to the tolerizing agent has ceased. Operationally, tolerance can be demonstrated by the lack of a T cell response upon reexposure to specific antigen in the absence of the tolerizing agent.

Down regulating or preventing one or more antigen functions (including without
25 limitation B lymphocyte antigen functions (such as , for example, B7)), *e.g.*, preventing high level lymphokine synthesis by activated T cells, will be useful in situations of tissue, skin and organ transplantation and in graft-versus-host disease (GVHD). For example, blockage of T cell function should result in reduced tissue destruction in tissue transplantation. Typically, in tissue transplants, rejection of the transplant is initiated through its recognition as foreign
30 by T cells, followed by an immune reaction that destroys the transplant. The administration of a molecule which inhibits or blocks interaction of a B7 lymphocyte antigen with its natural ligand(s) on immune cells (such as a soluble, monomeric form of a peptide having B7-2 activity alone or in conjunction with a monomeric form of a peptide having an activity of another B lymphocyte antigen (*e.g.*, B7-1, B7-3) or blocking antibody), prior to transplantation
35 can lead to the binding of the molecule to the natural ligand(s) on the immune cells without

transmitting the corresponding costimulatory signal. Blocking B lymphocyte antigen function in this matter prevents cytokine synthesis by immune cells, such as T cells, and thus acts as an immunosuppressant. Moreover, the lack of costimulation may also be sufficient to anergize the T cells, thereby inducing tolerance in a subject. Induction of long-term tolerance by B lymphocyte antigen-blocking reagents may avoid the necessity of repeated administration of these blocking reagents. To achieve sufficient immunosuppression or tolerance in a subject, it may also be necessary to block the function of a combination of B lymphocyte antigens.

The efficacy of particular blocking reagents in preventing organ transplant rejection or GVHD can be assessed using animal models that are predictive of efficacy in humans. Examples of appropriate systems which can be used include allogeneic cardiac grafts in rats and xenogeneic pancreatic islet cell grafts in mice, both of which have been used to examine the immunosuppressive effects of CTLA4Ig fusion proteins *in vivo* as described in Lenschow *et al.*, Science 257:789-792 (1992) and Turka *et al.*, Proc. Natl. Acad. Sci USA, 89:11102-11105 (1992). In addition, murine models of GVHD (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 846-847) can be used to determine the effect of blocking B lymphocyte antigen function *in vivo* on the development of that disease.

Blocking antigen function may also be therapeutically useful for treating autoimmune diseases. Many autoimmune disorders are the result of inappropriate activation of T cells that are reactive against self tissue and which promote the production of cytokines and autoantibodies involved in the pathology of the diseases. Preventing the activation of autoreactive T cells may reduce or eliminate disease symptoms. Administration of reagents which block costimulation of T cells by disrupting receptor:ligand interactions of B lymphocyte antigens can be used to inhibit T cell activation and prevent production of autoantibodies or T cell-derived cytokines which may be involved in the disease process. Additionally, blocking reagents may induce antigen-specific tolerance of autoreactive T cells which could lead to long-term relief from the disease. The efficacy of blocking reagents in preventing or alleviating autoimmune disorders can be determined using a number of well-characterized animal models of human autoimmune diseases. Examples include murine experimental autoimmune encephalitis, systemic lupus erythematosus in MRL/lpr/lpr mice or NZB hybrid mice, murine autoimmune collagen arthritis, diabetes mellitus in NOD mice and BB rats, and murine experimental myasthenia gravis (see Paul ed., Fundamental Immunology, Raven Press, New York, 1989, pp. 840-856).

Upregulation of an antigen function (preferably a B lymphocyte antigen function), as a means of up regulating immune responses, may also be useful in therapy. Upregulation of immune responses may be in the form of enhancing an existing immune response or eliciting

an initial immune response. For example, enhancing an immune response through stimulating B lymphocyte antigen function may be useful in cases of viral infection. In addition, systemic viral diseases such as influenza, the common cold, and encephalitis might be alleviated by the administration of stimulatory forms of B lymphocyte antigens systemically.

5 Alternatively, anti-viral immune responses may be enhanced in an infected patient by removing T cells from the patient, costimulating the T cells *in vitro* with viral antigen-pulsed APCs either expressing a peptide of the present invention or together with a stimulatory form of a soluble peptide of the present invention and reintroducing the *in vitro* activated T cells into the patient. Another method of enhancing anti-viral immune responses would be to isolate
10 infected cells from a patient, transfect them with a nucleic acid encoding a protein of the present invention as described herein such that the cells express all or a portion of the protein on their surface, and reintroduce the transfected cells into the patient. The infected cells would now be capable of delivering a costimulatory signal to, and thereby activate, T cells *in vivo*.

 In another application, up regulation or enhancement of antigen function (preferably
15 B lymphocyte antigen function) may be useful in the induction of tumor immunity. Tumor cells (*e.g.*, sarcoma, melanoma, lymphoma, leukemia, neuroblastoma, carcinoma) transfected with a nucleic acid encoding at least one peptide of the present invention can be administered to a subject to overcome tumor-specific tolerance in the subject. If desired, the tumor cell can be transfected to express a combination of peptides. For example, tumor cells obtained from
20 a patient can be transfected *ex vivo* with an expression vector directing the expression of a peptide having B7-2-like activity alone, or in conjunction with a peptide having B7-1-like activity and/or B7-3-like activity. The transfected tumor cells are returned to the patient to result in expression of the peptides on the surface of the transfected cell. Alternatively, gene therapy techniques can be used to target a tumor cell for transfection *in vivo*.

25 The presence of the peptide of the present invention having the activity of a B lymphocyte antigen(s) on the surface of the tumor cell provides the necessary costimulation signal to T cells to induce a T cell mediated immune response against the transfected tumor cells. In addition, tumor cells which lack MHC class I or MHC class II molecules, or which fail to reexpress sufficient amounts of MHC class I or MHC class II molecules, can be
30 transfected with nucleic acid encoding all or a portion of (*e.g.*, a cytoplasmic-domain truncated portion) of an MHC class I α chain protein and β_2 microglobulin protein or an MHC class II α chain protein and an MHC class II β chain protein to thereby express MHC class I or MHC class II proteins on the cell surface. Expression of the appropriate class I or class II MHC in conjunction with a peptide having the activity of a B lymphocyte antigen (*e.g.*, B7-1, B7-2, B7-
35 3) induces a T cell mediated immune response against the transfected tumor cell. Optionally,

a gene encoding an antisense construct which blocks expression of an MHC class II associated protein, such as the invariant chain, can also be cotransfected with a DNA encoding a peptide having the activity of a B lymphocyte antigen to promote presentation of tumor associated antigens and induce tumor specific immunity. Thus, the induction of a T cell mediated
5 immune response in a human subject may be sufficient to overcome tumor-specific tolerance in the subject.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for thymocyte or splenocyte cytotoxicity include, without limitation,
10 those described in: *Current Protocols in Immunology*, Ed by J. E. Coligan, A.M. Kruisbeek, D.H. Margulies, E.M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, *In Vitro* assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, *Immunologic studies in Humans*); Herrmann et al., *Proc. Natl. Acad. Sci. USA* 78:2488-2492, 1981; Herrmann et al., *J. Immunol.* 128:1968-1974, 1982; Handa et al., *J. Immunol.*
15 135:1564-1572, 1985; Takai et al., *J. Immunol.* 137:3494-3500, 1986; Takai et al., *J. Immunol.* 140:508-512, 1988; Herrmann et al., *Proc. Natl. Acad. Sci. USA* 78:2488-2492, 1981; Herrmann et al., *J. Immunol.* 128:1968-1974, 1982; Handa et al., *J. Immunol.* 135:1564-1572, 1985; Takai et al., *J. Immunol.* 137:3494-3500, 1986; Bowman et al., *J. Virology* 61:1992-1998; Takai et al., *J. Immunol.* 140:508-512, 1988; Bertagnoli et al.,
20 *Cellular Immunology* 133:327-341, 1991; Brown et al., *J. Immunol.* 153:3079-3092, 1994.

Assays for T-cell-dependent immunoglobulin responses and isotype switching (which will identify, among others, proteins that modulate T-cell dependent antibody responses and that affect Th1/Th2 profiles) include, without limitation, those described in: Maliszewski, *J. Immunol.* 144:3028-3033, 1990; and Assays for B cell function: *In vitro* antibody production,
25 Mond, J.J. and Brunswick, M. In *Current Protocols in Immunology*. J.E.e.a. Coligan eds. Vol 1 pp. 3.8.1-3.8.16, John Wiley and Sons, Toronto. 1994.

Mixed lymphocyte reaction (MLR) assays (which will identify, among others, proteins that generate predominantly Th1 and CTL responses) include, without limitation, those described in: *Current Protocols in Immunology*, Ed by J. E. Coligan, A.M. Kruisbeck, D.H.
30 Margulies, E.M. Shevach, W Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 3, *In Vitro* assays for Mouse Lymphocyte Function 3.1-3.19; Chapter 7, *Immunologic studies in Humans*); Takai et al., *J. Immunol.* 137:3494-3500, 1986; Takai et al., *J. Immunol.* 140:508-512, 1988; Bertagnoli et al., *J. Immunol.* 149:3778-3783, 1992.

Dendritic cell-dependent assays (which will identify, among others, proteins expressed
35 by dendritic cells that activate naive T-cells) include, without limitation, those described in:

Guery et al., J. Immunol. 134:536-544, 1995; Inaba et al., Journal of Experimental Medicine 173:549-559, 1991; Macatonia et al., Journal of Immunology 154:5071-5079, 1995; Porgador et al., Journal of Experimental Medicine 182:255-260, 1995; Nair et al., Journal of Virology 67:4062-4069, 1993; Huang et al., Science 264:961-965, 1994; Macatonia et al., Journal of
5 Experimental Medicine 169:1255-1264, 1989; Bhardwaj et al., Journal of Clinical Investigation 94:797-807, 1994; and Inaba et al., Journal of Experimental Medicine 172:631-640, 1990.

Assays for lymphocyte survival/apoptosis (which will identify, among others, proteins that prevent apoptosis after superantigen induction and proteins that regulate lymphocyte
10 homeostasis) include, without limitation, those described in: Darzynkiewicz et al., Cytometry 13:795-808, 1992; Gorczyca et al., Leukemia 7:659-670, 1993; Gorczyca et al., Cancer Research 53:1945-1951, 1993; Itoh et al., Cell 66:233-243, 1991; Zacharchuk, Journal of Immunology 145:4037-4045, 1990; Zamai et al., Cytometry 14:891-897, 1993; Gorczyca et al., International Journal of Oncology 1:639-648, 1992.

15 Assays for proteins that influence early steps of T-cell commitment and development include, without limitation, those described in: Antica et al., Blood 84:111-117, 1994; Fine et al., Cellular Immunology 155:111-122, 1994; Galy et al., Blood 85:2770-2778, 1995; Toki et al., Proc. Nat. Acad Sci. USA 88:7548-7551, 1991.

20 Hematopoiesis Regulating Activity

A protein of the present invention may be useful in regulation of hematopoiesis and, consequently, in the treatment of myeloid or lymphoid cell deficiencies. Even marginal biological activity in support of colony forming cells or of factor-dependent cell lines indicates involvement in regulating hematopoiesis, e.g. in supporting the growth and proliferation of
25 erythroid progenitor cells alone or in combination with other cytokines, thereby indicating utility, for example, in treating various anemias or for use in conjunction with irradiation/chemotherapy to stimulate the production of erythroid precursors and/or erythroid cells; in supporting the growth and proliferation of myeloid cells such as granulocytes and monocytes/macrophages (i.e., traditional CSF activity) useful, for example, in conjunction with
30 chemotherapy to prevent or treat consequent myelo-suppression; in supporting the growth and proliferation of megakaryocytes and consequently of platelets thereby allowing prevention or treatment of various platelet disorders such as thrombocytopenia, and generally for use in place of or complimentary to platelet transfusions; and/or in supporting the growth and proliferation of hematopoietic stem cells which are capable of maturing to any and all of the above-
35 mentioned hematopoietic cells and therefore find therapeutic utility in various stem cell

disorders (such as those usually treated with transplantation, including, without limitation, aplastic anemia and paroxysmal nocturnal hemoglobinuria), as well as in repopulating the stem cell compartment post irradiation/chemotherapy, either *in-vivo* or *ex-vivo* (i.e., in conjunction with bone marrow transplantation or with peripheral progenitor cell transplantation
5 (homologous or heterologous)) as normal cells or genetically manipulated for gene therapy.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Suitable assays for proliferation and differentiation of various hematopoietic lines are cited above.

10 Assays for embryonic stem cell differentiation (which will identify, among others, proteins that influence embryonic differentiation hematopoiesis) include, without limitation, those described in: Johansson et al. *Cellular Biology* 15:141-151, 1995; Keller et al., *Molecular and Cellular Biology* 13:473-486, 1993; McClanahan et al., *Blood* 81:2903-2915, 1993.

Assays for stem cell survival and differentiation (which will identify, among others,
15 proteins that regulate lympho-hematopoiesis) include, without limitation, those described in: Methylcellulose colony forming assays, Freshney, M.G. In *Culture of Hematopoietic Cells*. R.I. Freshney, et al. eds. Vol pp. 265-268, Wiley-Liss, Inc., New York, NY. 1994; Hirayama et al., *Proc. Natl. Acad. Sci. USA* 89:5907-5911, 1992; Primitive hematopoietic colony forming cells with high proliferative potential, McNiece, I.K. and Briddell, R.A. In *Culture of*
20 *Hematopoietic Cells*. R.I. Freshney, et al. eds. Vol pp. 23-39, Wiley-Liss, Inc., New York, NY. 1994; Neben et al., *Experimental Hematology* 22:353-359, 1994; Cobblestone area forming cell assay, Ploemacher, R.E. In *Culture of Hematopoietic Cells*. R.I. Freshney, et al. eds. Vol pp. 1-21, Wiley-Liss, Inc., New York, NY. 1994; Long term bone marrow cultures in the presence of stromal cells, Spooncer, E., Dexter, M. and Allen, T. In *Culture of*
25 *Hematopoietic Cells*. R.I. Freshney, et al. eds. Vol pp. 163-179, Wiley-Liss, Inc., New York, NY. 1994; Long term culture initiating cell assay, Sutherland, H.J. In *Culture of Hematopoietic Cells*. R.I. Freshney, et al. eds. Vol pp. 139-162, Wiley-Liss, Inc., New York, NY. 1994.

Tissue Growth Activity

30 A protein of the present invention also may have utility in compositions used for bone, cartilage, tendon, ligament and/or nerve tissue growth or regeneration, as well as for wound healing and tissue repair and replacement, and in the treatment of burns, incisions and ulcers.

A protein of the present invention, which induces cartilage and/or bone growth in circumstances where bone is not normally formed, has application in the healing of bone
35 fractures and cartilage damage or defects in humans and other animals. Such a preparation

employing a protein of the invention may have prophylactic use in closed as well as open fracture reduction and also in the improved fixation of artificial joints. *De novo* bone formation induced by an osteogenic agent contributes to the repair of congenital, trauma induced, or oncologic resection induced craniofacial defects, and also is useful in cosmetic plastic surgery.

A protein of this invention may also be used in the treatment of periodontal disease, and in other tooth repair processes. Such agents may provide an environment to attract bone-forming cells, stimulate growth of bone-forming cells or induce differentiation of progenitors of bone-forming cells. A protein of the invention may also be useful in the treatment of osteoporosis or osteoarthritis, such as through stimulation of bone and/or cartilage repair or by blocking inflammation or processes of tissue destruction (collagenase activity, osteoclast activity, etc.) mediated by inflammatory processes.

Another category of tissue regeneration activity that may be attributable to the protein of the present invention is tendon/ligament formation. A protein of the present invention, which induces tendon/ligament-like tissue or other tissue formation in circumstances where such tissue is not normally formed, has application in the healing of tendon or ligament tears, deformities and other tendon or ligament defects in humans and other animals. Such a preparation employing a tendon/ligament-like tissue inducing protein may have prophylactic use in preventing damage to tendon or ligament tissue, as well as use in the improved fixation of tendon or ligament to bone or other tissues, and in repairing defects to tendon or ligament tissue. *De novo* tendon/ligament-like tissue formation induced by a composition of the present invention contributes to the repair of congenital, trauma induced, or other tendon or ligament defects of other origin, and is also useful in cosmetic plastic surgery for attachment or repair of tendons or ligaments. The compositions of the present invention may provide an environment to attract tendon- or ligament-forming cells, stimulate growth of tendon- or ligament-forming cells, induce differentiation of progenitors of tendon- or ligament-forming cells, or induce growth of tendon/ligament cells or progenitors *ex vivo* for return *in vivo* to effect tissue repair. The compositions of the invention may also be useful in the treatment of tendinitis, carpal tunnel syndrome and other tendon or ligament defects. The compositions may also include an appropriate matrix and/or sequestering agent as a carrier as is well known in the art.

The protein of the present invention may also be useful for proliferation of neural cells and for regeneration of nerve and brain tissue, *i.e.* for the treatment of central and peripheral nervous system diseases and neuropathies, as well as mechanical and traumatic disorders, which involve degeneration, death or trauma to neural cells or nerve tissue. More specifically,

a protein may be used in the treatment of diseases of the peripheral nervous system, such as peripheral nerve injuries, peripheral neuropathy and localized neuropathies, and central nervous system diseases, such as Alzheimer's, Parkinson's disease, Huntington's disease, amyotrophic lateral sclerosis, and Shy-Drager syndrome. Further conditions which may be
5 treated in accordance with the present invention include mechanical and traumatic disorders, such as spinal cord disorders, head trauma and cerebrovascular diseases such as stroke. Peripheral neuropathies resulting from chemotherapy or other medical therapies may also be treatable using a protein of the invention.

Proteins of the invention may also be useful to promote better or faster closure of non-
10 healing wounds, including without limitation pressure ulcers, ulcers associated with vascular insufficiency, surgical and traumatic wounds, and the like.

It is expected that a protein of the present invention may also exhibit activity for generation or regeneration of other tissues, such as organs (including, for example, pancreas, liver, intestine, kidney, skin, endothelium), muscle (smooth, skeletal or cardiac) and vascular
15 (including vascular endothelium) tissue, or for promoting the growth of cells comprising such tissues. Part of the desired effects may be by inhibition or modulation of fibrotic scarring to allow normal tissue to regenerate. A protein of the invention may also exhibit angiogenic activity.

A protein of the present invention may also be useful for gut protection or regeneration
20 and treatment of lung or liver fibrosis, reperfusion injury in various tissues, and conditions resulting from systemic cytokine damage.

A protein of the present invention may also be useful for promoting or inhibiting differentiation of tissues described above from precursor tissues or cells; or for inhibiting the growth of tissues described above.

25 The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for tissue generation activity include, without limitation, those described in: International Patent Publication No. WO95/16035 (bone, cartilage, tendon); International Patent Publication No. WO95/05846 (nerve, neuronal); International Patent Publication No.
30 WO91/07491 (skin, endothelium).

Assays for wound healing activity include, without limitation, those described in: Winter, Epidermal Wound Healing, pps. 71-112 (Maibach, HI and Rovee, DT, eds.), Year Book Medical Publishers, Inc., Chicago, as modified by Eaglstein and Mertz, J. Invest. Dermatol 71:382-84 (1978).

Activin/Inhibin Activity

A protein of the present invention may also exhibit activin- or inhibin-related activities. Inhibins are characterized by their ability to inhibit the release of follicle stimulating hormone (FSH), while activins are characterized by their ability to stimulate the release of follicle stimulating hormone (FSH). Thus, a protein of the present invention, alone or in heterodimers with a member of the inhibin α family, may be useful as a contraceptive based on the ability of inhibins to decrease fertility in female mammals and decrease spermatogenesis in male mammals. Administration of sufficient amounts of other inhibins can induce infertility in these mammals. Alternatively, the protein of the invention, as a homodimer or as a heterodimer with other protein subunits of the inhibin- β group, may be useful as a fertility inducing therapeutic, based upon the ability of activin molecules in stimulating FSH release from cells of the anterior pituitary. See, for example, United States Patent 4,798,885. A protein of the invention may also be useful for advancement of the onset of fertility in sexually immature mammals, so as to increase the lifetime reproductive performance of domestic animals such as cows, sheep and pigs.

The activity of a protein of the invention may, among other means, be measured by the following methods:

Assays for activin/inhibin activity include, without limitation, those described in: Vale et al., Endocrinology 91:562-572, 1972; Ling et al., Nature 321:779-782, 1986; Vale et al., Nature 321:776-779, 1986; Mason et al., Nature 318:659-663, 1985; Forage et al., Proc. Natl. Acad. Sci. USA 83:3091-3095, 1986.

Chemotactic/Chemokinetic Activity

A protein of the present invention may have chemotactic or chemokinetic activity (e.g., act as a chemokine) for mammalian cells, including, for example, monocytes, fibroblasts, neutrophils, T-cells, mast cells, eosinophils, epithelial and/or endothelial cells. Chemotactic and chemokinetic proteins can be used to mobilize or attract a desired cell population to a desired site of action. Chemotactic or chemokinetic proteins provide particular advantages in treatment of wounds and other trauma to tissues, as well as in treatment of localized infections. For example, attraction of lymphocytes, monocytes or neutrophils to tumors or sites of infection may result in improved immune responses against the tumor or infecting agent.

A protein or peptide has chemotactic activity for a particular cell population if it can stimulate, directly or indirectly, the directed orientation or movement of such cell population. Preferably, the protein or peptide has the ability to directly stimulate directed movement of cells. Whether a particular protein has chemotactic activity for a population of cells can be

readily determined by employing such protein or peptide in any known assay for cell chemotaxis.

The activity of a protein of the invention may, among other means, be measured by the following methods:

- 5 Assays for chemotactic activity (which will identify proteins that induce or prevent chemotaxis) consist of assays that measure the ability of a protein to induce the migration of cells across a membrane as well as the ability of a protein to induce the adhesion of one cell population to another cell population. Suitable assays for movement and adhesion include, without limitation, those described in: Current Protocols in Immunology, Ed by J.E. Coligan, 10 A.M. Kruisbeek, D.H. Margulies, E.M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 6.12, Measurement of alpha and beta Chemokines 6.12.1-6.12.28; Taub et al. J. Clin. Invest. 95:1370-1376, 1995; Lind et al. APMIS 103:140-146, 1995; Muller et al Eur. J. Immunol. 25: 1744-1748; Gruber et al. J. of Immunol. 152:5860-5867, 1994; Johnston et al. J. of Immunol. 153: 1762-1768, 1994.

15

Hemostatic and Thrombolytic Activity

- A protein of the invention may also exhibit hemostatic or thrombolytic activity. As a result, such a protein is expected to be useful in treatment of various coagulation disorders (including hereditary disorders, such as hemophilias) or to enhance coagulation and other 20 hemostatic events in treating wounds resulting from trauma, surgery or other causes. A protein of the invention may also be useful for dissolving or inhibiting formation of thromboses and for treatment and prevention of conditions resulting therefrom (such as, for example, infarction of cardiac and central nervous system vessels (e.g., stroke).

- 25 The activity of a protein of the invention may, among other means, be measured by the following methods:

Assay for hemostatic and thrombolytic activity include, without limitation, those described in: Linet et al., J. Clin. Pharmacol. 26:131-140, 1986; Burdick et al., Thrombosis Res. 45:413-419, 1987; Humphrey et al., Fibrinolysis 5:71-79 (1991); Schaub, Prostaglandins 35:467-474, 1988.

30

Receptor/Ligand Activity

- A protein of the present invention may also demonstrate activity as receptors, receptor ligands or inhibitors or agonists of receptor/ligand interactions. Examples of such receptors and ligands include, without limitation, cytokine receptors and their ligands, receptor kinases 35 and their ligands, receptor phosphatases and their ligands, receptors involved in cell-cell

interactions and their ligands (including without limitation, cellular adhesion molecules (such as selectins, integrins and their ligands) and receptor/ligand pairs involved in antigen presentation, antigen recognition and development of cellular and humoral immune responses). Receptors and ligands are also useful for screening of potential peptide or small molecule
5 inhibitors of the relevant receptor/ligand interaction. A protein of the present invention (including, without limitation, fragments of receptors and ligands) may themselves be useful as inhibitors of receptor/ligand interactions.

The activity of a protein of the invention may, among other means, be measured by the following methods:

10 Suitable assays for receptor-ligand activity include without limitation those described in: Current Protocols in Immunology, Ed by J.E. Coligan, A.M. Kruisbeek, D.H. Margulies, E.M. Shevach, W. Strober, Pub. Greene Publishing Associates and Wiley-Interscience (Chapter 7.28, Measurement of Cellular Adhesion under static conditions 7.28.1-7.28.22), Takai et al., Proc. Natl. Acad. Sci. USA 84:6864-6868, 1987; Bierer et al., J. Exp. Med.
15 168:1145-1156, 1988; Rosenstein et al., J. Exp. Med. 169:149-160 1989; Stoltenborg et al., J. Immunol. Methods 175:59-68, 1994; Stitt et al., Cell 80:661-670, 1995.

Anti-Inflammatory Activity

Proteins of the present invention may also exhibit anti-inflammatory activity. The anti-
20 inflammatory activity may be achieved by providing a stimulus to cells involved in the inflammatory response, by inhibiting or promoting cell-cell interactions (such as, for example, cell adhesion), by inhibiting or promoting chemotaxis of cells involved in the inflammatory process, inhibiting or promoting cell extravasation, or by stimulating or suppressing production of other factors which more directly inhibit or promote an inflammatory response. Proteins
25 exhibiting such activities can be used to treat inflammatory conditions including chronic or acute conditions), including without limitation inflammation associated with infection (such as septic shock, sepsis or systemic inflammatory response syndrome (SIRS)), ischemia-reperfusion injury, endotoxin lethality, arthritis, complement-mediated hyperacute rejection, nephritis, cytokine or chemokine-induced lung injury, inflammatory bowel disease, Crohn's
30 disease or resulting from over production of cytokines such as TNF or IL-1. Proteins of the invention may also be useful to treat anaphylaxis and hypersensitivity to an antigenic substance or material.

Tumor Inhibition Activity

In addition to the activities described above for immunological treatment or prevention of tumors, a protein of the invention may exhibit other anti-tumor activities. A protein may inhibit tumor growth directly or indirectly (such as, for example, via ADCC). A protein may exhibit its tumor inhibitory activity by acting on tumor tissue or tumor precursor tissue, by
5 inhibiting formation of tissues necessary to support tumor growth (such as, for example, by inhibiting angiogenesis), by causing production of other factors, agents or cell types which inhibit tumor growth, or by suppressing, eliminating or inhibiting factors, agents or cell types which promote tumor growth.

10

Other Activities

A protein of the invention may also exhibit one or more of the following additional activities or effects: inhibiting the growth, infection or function of, or killing, infectious agents, including, without limitation, bacteria, viruses, fungi and other parasites; effecting (suppressing
15 or enhancing) bodily characteristics, including, without limitation, height, weight, hair color, eye color, skin, fat to lean ratio or other tissue pigmentation, or organ or body part size or shape (such as, for example, breast augmentation or diminution, change in bone form or shape); effecting biorhythms or circadian cycles or rhythms; effecting the fertility of male or female subjects; effecting the metabolism, catabolism, anabolism, processing, utilization, storage or
20 elimination of dietary fat, lipid, protein, carbohydrate, vitamins, minerals, cofactors or other nutritional factors or component(s); effecting behavioral characteristics, including, without limitation, appetite, libido, stress, cognition (including cognitive disorders), depression (including depressive disorders) and violent behaviors; providing analgesic effects or other pain reducing effects; promoting differentiation and growth of embryonic stem cells in lineages
25 other than hematopoietic lineages; hormonal or endocrine activity; in the case of enzymes, correcting deficiencies of the enzyme and treating deficiency-related diseases; treatment of hyperproliferative disorders (such as, for example, psoriasis); immunoglobulin-like activity (such as, for example, the ability to bind antigens or complement); and the ability to act as an antigen in a vaccine composition to raise an immune response against such protein or another
30 material or entity which is cross-reactive with such protein.

ADMINISTRATION AND DOSING

A protein of the present invention (from whatever source derived, including without
35 limitation from recombinant and non-recombinant sources) may be used in a pharmaceutical

composition when combined with a pharmaceutically acceptable carrier. Such a composition may also contain (in addition to protein and a carrier) diluents, fillers, salts, buffers, stabilizers, solubilizers, and other materials well known in the art. The term "pharmaceutically acceptable" means a non-toxic material that does not interfere with the effectiveness of the biological activity of the active ingredient(s). The characteristics of the carrier will depend on the route of administration. The pharmaceutical composition of the invention may also contain cytokines, lymphokines, or other hematopoietic factors such as M-CSF, GM-CSF, TNF, IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-11, IL-12, IL-13, IL-14, IL-15, IFN, TNF0, TNF1, TNF2, G-CSF, Meg-CSF, thrombopoietin, stem cell factor, and erythropoietin.

10 The pharmaceutical composition may further contain other agents which either enhance the activity of the protein or compliment its activity or use in treatment. Such additional factors and/or agents may be included in the pharmaceutical composition to produce a synergistic effect with protein of the invention, or to minimize side effects. Conversely, protein of the present invention may be included in formulations of the particular cytokine, lymphokine, other

15 hematopoietic factor, thrombolytic or anti-thrombotic factor, or anti-inflammatory agent to minimize side effects of the cytokine, lymphokine, other hematopoietic factor, thrombolytic or anti-thrombotic factor, or anti-inflammatory agent.

A protein of the present invention may be active in multimers (e.g., heterodimers or homodimers) or complexes with itself or other proteins. As a result, pharmaceutical compositions of the invention may comprise a protein of the invention in such multimeric or

20 complexed form.

The pharmaceutical composition of the invention may be in the form of a complex of the protein(s) of present invention along with protein or peptide antigens. The protein and/or peptide antigen will deliver a stimulatory signal to both B and T lymphocytes. B lymphocytes will respond to antigen through their surface immunoglobulin receptor. T lymphocytes will respond to antigen through the T cell receptor (TCR) following presentation of the antigen by MHC proteins. MHC and structurally related proteins including those encoded by class I and class II MHC genes on host cells will serve to present the peptide antigen(s) to T lymphocytes. The antigen components could also be supplied as purified MHC-peptide complexes alone or

25 with co-stimulatory molecules that can directly signal T cells. Alternatively antibodies able to bind surface immunoglobulin and other molecules on B cells as well as antibodies able to bind the TCR and other molecules on T cells can be combined with the pharmaceutical composition of the invention.

30

The pharmaceutical composition of the invention may be in the form of a liposome in which protein of the present invention is combined, in addition to other pharmaceutically

35

acceptable carriers, with amphipathic agents such as lipids which exist in aggregated form as micelles, insoluble monolayers, liquid crystals, or lamellar layers in aqueous solution. Suitable lipids for liposomal formulation include, without limitation, monoglycerides, diglycerides, sulfatides, lysolecithin, phospholipids, saponin, bile acids, and the like. Preparation of such liposomal formulations is within the level of skill in the art, as disclosed, for example, in U.S. Patent No. 4,235,871; U.S. Patent No. 4,501,728; U.S. Patent No. 4,837,028; and U.S. Patent No. 4,737,323, all of which are incorporated herein by reference.

As used herein, the term "therapeutically effective amount" means the total amount of each active component of the pharmaceutical composition or method that is sufficient to show a meaningful patient benefit, i.e., treatment, healing, prevention or amelioration of the relevant medical condition, or an increase in rate of treatment, healing, prevention or amelioration of such conditions. When applied to an individual active ingredient, administered alone, the term refers to that ingredient alone. When applied to a combination, the term refers to combined amounts of the active ingredients that result in the therapeutic effect, whether administered in combination, serially or simultaneously.

In practicing the method of treatment or use of the present invention, a therapeutically effective amount of protein of the present invention is administered to a mammal having a condition to be treated. Protein of the present invention may be administered in accordance with the method of the invention either alone or in combination with other therapies such as treatments employing cytokines, lymphokines or other hematopoietic factors. When co-administered with one or more cytokines, lymphokines or other hematopoietic factors, protein of the present invention may be administered either simultaneously with the cytokine(s), lymphokine(s), other hematopoietic factor(s), thrombolytic or anti-thrombotic factors, or sequentially. If administered sequentially, the attending physician will decide on the appropriate sequence of administering protein of the present invention in combination with cytokine(s), lymphokine(s), other hematopoietic factor(s), thrombolytic or anti-thrombotic factors.

Administration of protein of the present invention used in the pharmaceutical composition or to practice the method of the present invention can be carried out in a variety of conventional ways, such as oral ingestion, inhalation, topical application or cutaneous, subcutaneous, intraperitoneal, parenteral or intravenous injection. Intravenous administration to the patient is preferred.

When a therapeutically effective amount of protein of the present invention is administered orally, protein of the present invention will be in the form of a tablet, capsule, powder, solution or elixir. When administered in tablet form, the pharmaceutical composition

of the invention may additionally contain a solid carrier such as a gelatin or an adjuvant. The tablet, capsule, and powder contain from about 5 to 95% protein of the present invention, and preferably from about 25 to 90% protein of the present invention. When administered in liquid form, a liquid carrier such as water, petroleum, oils of animal or plant origin such as peanut oil, mineral oil, soybean oil, or sesame oil, or synthetic oils may be added. The liquid form of the pharmaceutical composition may further contain physiological saline solution, dextrose or other saccharide solution, or glycols such as ethylene glycol, propylene glycol or polyethylene glycol. When administered in liquid form, the pharmaceutical composition contains from about 0.5 to 90% by weight of protein of the present invention, and preferably from about 1 to 50% protein of the present invention.

When a therapeutically effective amount of protein of the present invention is administered by intravenous, cutaneous or subcutaneous injection, protein of the present invention will be in the form of a pyrogen-free, parenterally acceptable aqueous solution. The preparation of such parenterally acceptable protein solutions, having due regard to pH, isotonicity, stability, and the like, is within the skill in the art. A preferred pharmaceutical composition for intravenous, cutaneous, or subcutaneous injection should contain, in addition to protein of the present invention, an isotonic vehicle such as Sodium Chloride Injection, Ringer's Injection, Dextrose Injection, Dextrose and Sodium Chloride Injection, Lactated Ringer's Injection, or other vehicle as known in the art. The pharmaceutical composition of the present invention may also contain stabilizers, preservatives, buffers, antioxidants, or other additives known to those of skill in the art.

The amount of protein of the present invention in the pharmaceutical composition of the present invention will depend upon the nature and severity of the condition being treated, and on the nature of prior treatments which the patient has undergone. Ultimately, the attending physician will decide the amount of protein of the present invention with which to treat each individual patient. Initially, the attending physician will administer low doses of protein of the present invention and observe the patient's response. Larger doses of protein of the present invention may be administered until the optimal therapeutic effect is obtained for the patient, and at that point the dosage is not increased further. It is contemplated that the various pharmaceutical compositions used to practice the method of the present invention should contain about 0.01 μ g to about 100 mg (preferably about 0.1 μ g to about 10 mg, more preferably about 0.1 μ g to about 1 mg) of protein of the present invention per kg body weight.

The duration of intravenous therapy using the pharmaceutical composition of the present invention will vary, depending on the severity of the disease being treated and the condition and potential idiosyncratic response of each individual patient. It is contemplated

that the duration of each application of the protein of the present invention will be in the range of 12 to 24 hours of continuous intravenous administration. Ultimately the attending physician will decide on the appropriate duration of intravenous therapy using the pharmaceutical composition of the present invention.

5 Protein of the invention may also be used to immunize animals to obtain polyclonal and monoclonal antibodies which specifically react with the protein. Such antibodies may be obtained using either the entire protein or fragments thereof as an immunogen. The peptide immunogens additionally may contain a cysteine residue at the carboxyl terminus, and are conjugated to a hapten such as keyhole limpet hemocyanin (KLH). Methods for synthesizing
10 such peptides are known in the art, for example, as in R.P. Merrifield, J. Amer.Chem.Soc. 85, 2149-2154 (1963); J.L. Krstenansky, *et al.*, FEBS Lett. 211, 10 (1987). Monoclonal antibodies binding to the protein of the invention may be useful diagnostic agents for the immunodetection of the protein. Neutralizing monoclonal antibodies binding to the protein may also be useful therapeutics for both conditions associated with the protein and also in the
15 treatment of some forms of cancer where abnormal expression of the protein is involved. In the case of cancerous cells or leukemic cells, neutralizing monoclonal antibodies against the protein may be useful in detecting and preventing the metastatic spread of the cancerous cells, which may be mediated by the protein.

For compositions of the present invention which are useful for bone, cartilage, tendon
20 or ligament regeneration, the therapeutic method includes administering the composition topically, systematically, or locally as an implant or device. When administered, the therapeutic composition for use in this invention is, of course, in a pyrogen-free, physiologically acceptable form. Further, the composition may desirably be encapsulated or injected in a viscous form for delivery to the site of bone, cartilage or tissue damage. Topical
25 administration may be suitable for wound healing and tissue repair. Therapeutically useful agents other than a protein of the invention which may also optionally be included in the composition as described above, may alternatively or additionally, be administered simultaneously or sequentially with the composition in the methods of the invention. Preferably for bone and/or cartilage formation, the composition would include a matrix capable
30 of delivering the protein-containing composition to the site of bone and/or cartilage damage, providing a structure for the developing bone and cartilage and optimally capable of being resorbed into the body. Such matrices may be formed of materials presently in use for other implanted medical applications.

The choice of matrix material is based on biocompatibility, biodegradability,
35 mechanical properties, cosmetic appearance and interface properties. The particular

application of the compositions will define the appropriate formulation. Potential matrices for the compositions may be biodegradable and chemically defined calcium sulfate, tricalciumphosphate, hydroxyapatite, polylactic acid, polyglycolic acid and polyanhydrides. Other potential materials are biodegradable and biologically well-defined, such as bone or dermal collagen. Further matrices are comprised of pure proteins or extracellular matrix components. Other potential matrices are nonbiodegradable and chemically defined, such as sintered hydroxapatite, bioglass, aluminates, or other ceramics. Matrices may be comprised of combinations of any of the above mentioned types of material, such as polylactic acid and hydroxyapatite or collagen and tricalciumphosphate. The bioceramics may be altered in composition, such as in calcium-aluminate-phosphate and processing to alter pore size, particle size, particle shape, and biodegradability.

Presently preferred is a 50:50 (mole weight) copolymer of lactic acid and glycolic acid in the form of porous particles having diameters ranging from 150 to 800 microns. In some applications, it will be useful to utilize a sequestering agent, such as carboxymethyl cellulose or autologous blood clot, to prevent the protein compositions from disassociating from the matrix.

A preferred family of sequestering agents is cellulosic materials such as alkylcelluloses (including hydroxyalkylcelluloses), including methylcellulose, ethylcellulose, hydroxyethylcellulose, hydroxypropylcellulose, hydroxypropyl-methylcellulose, and carboxymethylcellulose, the most preferred being cationic salts of carboxymethylcellulose (CMC). Other preferred sequestering agents include hyaluronic acid, sodium alginate, poly(ethylene glycol), polyoxyethylene oxide, carboxyvinyl polymer and poly(vinyl alcohol). The amount of sequestering agent useful herein is 0.5-20 wt%, preferably 1-10 wt% based on total formulation weight, which represents the amount necessary to prevent desorption of the protein from the polymer matrix and to provide appropriate handling of the composition, yet not so much that the progenitor cells are prevented from infiltrating the matrix, thereby providing the protein the opportunity to assist the osteogenic activity of the progenitor cells.

In further compositions, proteins of the invention may be combined with other agents beneficial to the treatment of the bone and/or cartilage defect, wound, or tissue in question. These agents include various growth factors such as epidermal growth factor (EGF), platelet derived growth factor (PDGF), transforming growth factors (TGF- α and TGF- β), and insulin-like growth factor (IGF).

The therapeutic compositions are also presently valuable for veterinary applications. Particularly domestic animals and thoroughbred horses, in addition to humans, are desired patients for such treatment with proteins of the present invention.

The dosage regimen of a protein-containing pharmaceutical composition to be used in tissue regeneration will be determined by the attending physician considering various factors which modify the action of the proteins, e.g., amount of tissue weight desired to be formed, the site of damage, the condition of the damaged tissue, the size of a wound, type of damaged tissue (e.g., bone), the patient's age, sex, and diet, the severity of any infection, time of administration and other clinical factors. The dosage may vary with the type of matrix used in the reconstitution and with inclusion of other proteins in the pharmaceutical composition. For example, the addition of other known growth factors, such as IGF I (insulin like growth factor I), to the final composition, may also effect the dosage. Progress can be monitored by periodic assessment of tissue/bone growth and/or repair, for example, X-rays, histomorphometric determinations and tetracycline labeling.

Polynucleotides of the present invention can also be used for gene therapy. Such polynucleotides can be introduced either *in vivo* or *ex vivo* into cells for expression in a mammalian subject. Polynucleotides of the invention may also be administered by other known methods for introduction of nucleic acid into a cell or organism (including, without limitation, in the form of viral vectors or naked DNA).

Cells may also be cultured *ex vivo* in the presence of proteins of the present invention in order to proliferate or to produce a desired effect on or activity in such cells. Treated cells can then be introduced *in vivo* for therapeutic purposes.

20

Patent and literature references cited herein are incorporated by reference as if fully set forth.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

- (i) APPLICANT: Jacobs, Kenneth
McCoy, John
LaVallie, Edward
Racie, Lisa
Merberg, David
Treacy, Maurice
Spaulding, Vikki
- (ii) TITLE OF INVENTION: SECRETED PROTEINS
- (iii) NUMBER OF SEQUENCES: 54
- (iv) CORRESPONDENCE ADDRESS:
 - (A) ADDRESSEE: Genetics Institute, Inc.
 - (B) STREET: 87 CambridgePark Drive
 - (C) CITY: Cambridge
 - (D) STATE: Massachusetts
 - (E) COUNTRY: U.S.A.
 - (F) ZIP: 02140
- (v) COMPUTER READABLE FORM:
 - (A) MEDIUM TYPE: Floppy disk
 - (B) COMPUTER: IBM PC compatible
 - (C) OPERATING SYSTEM: PC-DOS/MS-DOS
 - (D) SOFTWARE: PatentIn Release #1.0, Version #1.30
- (vi) CURRENT APPLICATION DATA:
 - (A) APPLICATION NUMBER:
 - (B) FILING DATE:
 - (C) CLASSIFICATION:
- (viii) ATTORNEY/AGENT INFORMATION:
 - (A) NAME: Brown, Scott A.
 - (B) REGISTRATION NUMBER: 32,724
- (ix) TELECOMMUNICATION INFORMATION:
 - (A) TELEPHONE: (617) 498-8224
 - (B) TELEFAX: (617) 876-5851

(2) INFORMATION FOR SEQ ID NO:1:

- (i) SEQUENCE CHARACTERISTICS:
 - (A) LENGTH: 276 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: double
 - (D) TOPOLOGY: linear
- (ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

```

AAGCTTGGGG TTTTCTGGGC TACTACGATG GCGATGAGTT TCGAGTGGCC GTGGCAGTAC      60
CGCTTCCCGC CCTTCTTTAC GTTACAGCCG AACGTGGACA CCCGGCAGAA GCAGCTGGCC      120
GCCTGGTGCT CTCTGGTTCT GTCCTTCTGC CGCCTGCACA AACAGTCCAG CATGACGGTG      180
ATGGAAGCCC AGGAGAGCCC GCTTTTCAAC AACGTCAAGC TACAGCGGAA ACTTCCTGTG      240
GAGTCAATTC AGATTGTATT AGAAGAACTG AGAAAG                                276

```

(2) INFORMATION FOR SEQ ID NO:2:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 83 amino acids
 - (B) TYPE: amino acid
 - (C) STRANDEDNESS:
 - (D) TOPOLOGY: linear

- (ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

```

Met Ala Met Ser Phe Glu Trp Pro Trp Gln Tyr Arg Phe Pro Pro Phe
1.          5          10          15

Phe Thr Leu Gln Pro Asn Val Asp Thr Arg Gln Lys Gln Leu Ala Ala
          20          25          30

Trp Cys Ser Leu Val Leu Ser Phe Cys Arg Leu His Lys Gln Ser Ser
          35          40          45

Met Thr Val Met Glu Ala Gln Glu Ser Pro Leu Phe Asn Asn Val Lys
          50          55          60

Leu Gln Arg Lys Leu Pro Val Glu Ser Ile Gln Ile Val Leu Glu Glu
65          70          75          80

Leu Arg Lys

```

(2) INFORMATION FOR SEQ ID NO:3:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 246 base pairs
 - (B) TYPE: nucleic acid
 - (C) STRANDEDNESS: double
 - (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

GTGAGTACAT ACACACANGC GCNTGCAGCA CANGATTNTG TCTCATCGTC NTCCCACCCN	60
NNNNGGNGNN GNTGCCTCCC TTAGTCAGGN GANGATGNAT CCTTTCCNAG GGGNTGGGGG	120
GNANCATTGG ATGCGGGCAG CNTTCCAGGC AANATGAAGA TNGGAGGCCC ACGGGCATGG	180
CAGTGAGAGG NGTGGCCCCA CACNGATTTA TGATNTTGAA ATCTCAACTC CAAAAAAGA	240
AAAAAA	246

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 632 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

AGCTTCGGAA TAATAATTTT GGCAAATCTA TCTTCTGAAC CACTCATTTT TGTGGTCTTA	60
ATGGCTCCAA TTTGGGGACC AATAATGTTC ATTGTCTCAG GATCCCTGTC AATTGCAGCA	120
GGAGTGAAAC CTACAAAAG CCTGATCATC AGCAGTCTAA CTCTGAACAC TATCACCTCT	180
GTGTTGGCTG CAACTGCAAG CATAATGGGT GTAGTCAGTG TGGCTGTGGG TTCACAGTTT	240
CCGTTTCGGT ATAATTATAC AATCACCAAG GGTTCGGATA TTTTGATGTT AATTTTAAAT	300
ATGCTAGAAT TCTGCATTGC TGTGTCCATC TCTGCTTTTG GATGTAAAGC TTCCTGTTGT	360
AACTCCAGCG AGGTTCTTGT AGTGCTACCA TCAAATCCTG CTGTGACTGT GATGGCACCC	420
CCCACACCAC TTAATGAAGG TTTGAGGCCA CAAAAGATC AACAGACAAA TGCTCCAGAA	480
ATCTATGCTG ACTGTGACAC AAGAAGCCTC ACATGAAGAA ATTACCAGTA TCCAACCTCG	540
ATACTGATAG ACTTGTGAT ATTATTATTA TATGTAATCC AATTATGAAC TGTGTGTGTA	600
TAGAGAGATA ATAAATTCAA AATTATGTTC TC	632

(2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 151 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

Met	Ala	Pro	Ile	Trp	Gly	Pro	Ile	Met	Phe	Ile	Val	Ser	Gly	Ser	Leu	1	5	10	15
Ser	Ile	Ala	Ala	Gly	Val	Lys	Pro	Thr	Lys	Ser	Leu	Ile	Ile	Ser	Ser	20	25	30	
Leu	Thr	Leu	Asn	Thr	Ile	Thr	Ser	Val	Leu	Ala	Ala	Thr	Ala	Ser	Ile	35	40	45	
Met	Gly	Val	Val	Ser	Val	Ala	Val	Gly	Ser	Gln	Phe	Pro	Phe	Arg	Tyr	50	55	60	
Asn	Tyr	Thr	Ile	Thr	Lys	Gly	Leu	Asp	Ile	Leu	Met	Leu	Ile	Leu	Asn	65	70	75	80
Met	Leu	Glu	Phe	Cys	Ile	Ala	Val	Ser	Ile	Ser	Ala	Phe	Gly	Cys	Lys	85	90	95	
Ala	Ser	Cys	Cys	Asn	Ser	Ser	Glu	Val	Leu	Val	Val	Leu	Pro	Ser	Asn	100	105	110	
Pro	Ala	Val	Thr	Val	Met	Ala	Pro	Pro	Thr	Pro	Leu	Asn	Glu	Gly	Leu	115	120	125	
Arg	Pro	Pro	Lys	Asp	Gln	Gln	Thr	Asn	Ala	Pro	Glu	Ile	Tyr	Ala	Asp	130	135	140	
Cys	Asp	Thr	Arg	Ser	Leu	Thr	145	150											

(2) INFORMATION FOR SEQ ID NO:6:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 365 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CTATGGGGAC CAAAGTGNTT TTTCNTTCAG GAAGTGGAGA TGCATGGCCA TCTCCCCCTC	60
CCTTTTTCCT TCTCNTGNTT TTCTTTCCCC ATAGAAAGTA CCTTGAAGTA GCACAGTCCG	120
TCCTTGCATG TGCNCGNGCT NTCNTTTGAG TAAAAGTATA CATGGAGTAA AAATCATATT	180
AAGCATCAGA TTCAACTTAT ATTTTNTATT TCATNTTCTT CCTTTCCCTT CTCCCACNTT	240
NTACTGGGCA TAATTATATN TTAATCATAT ATGGAAATGT GCAACATATG GTATTTGTTA	300
AATACGTTTG TTTTATTGTC AGAGCAAAAA TAAATCAAAT TAGAAGCAAA AAAAAAAAAA	360
AAAAA	365

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 689 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: double
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

CCCANAGAGN CCTAGGAAGA TGAACAAACG ACAGCTCTAC TACCAGGTTT TAACTTTG	60
CATGATCGTG TCTTCTGCGC TCATGATCTG GAAAGGCCTG ATTGTTCTCA CGGGCAGCGA	120
GAGTCCCATC GTGGWGGTAC TCAGTGGCAG TATGGAGCCG GCCTTCCACA GAGGAGATCT	180
BCTGTTCCCTC ACGAATTTCC GGGAGGACCC CATCAGAGCT GGTGAAATAG TTGTTTTTAA	240
GGTTGAAGGA AGAGACATTC CGATAGTTCA CAGAGTAATC AAGGTTCTG AAAAAAGATAA	300
TGGTGACATC AARTTTCTGA CTAAAGGAGA TAATAATGAA GTYGATGATA GAGGCTTGTA	360
CAAAGAAGGC CAGAACTGGC TGGAAAAGAA GGACGTGGTG GGAAGAGCAA GANGGTTTTT	420
ACCATATGTT GGTATGGTCA CCATAATAAT GAATGACTAT CCAAATTC AATATGCTCT	480
TTTGGCTGTA ATGGGTGCAT ATGTGTTACT AAAACGTGAA TCCTAAAATG AGAAGCAGTT	540
CCTGGGACCA GATTGAAATG AATTCTGTTG AAAAGAGAA AACTAATAT ATTTGAGATG	600
TTCCATTTTC TGTATAAAAAG GGAACAGTGT GGAGATGTTT TTGTCTTGTC CAAATAAAAG	660
ATTCACCAGT AAAAAAAAAA AAAAAAAAAA	689

(2) INFORMATION FOR SEQ ID NO:8:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 168 amino acids
 (B) TYPE: amino acid
 (C) STRANDEDNESS:
 (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

```

Met Asn Lys Arg Gln Leu Tyr Tyr Gln Val Leu Asn Phe Ala Met Ile
1           5           10           15

Val Ser Ser Ala Leu Met Ile Trp Lys Gly Leu Ile Val Leu Thr Gly
          20           25           30

Ser Glu Ser Pro Ile Val Xaa Val Leu Ser Gly Ser Met Glu Pro Ala
          35           40           45

Phe His Arg Gly Asp Leu Leu Phe Leu Thr Asn Phe Arg Glu Asp Pro
          50           55           60

Ile Arg Ala Gly Glu Ile Val Val Phe Lys Val Glu Gly Arg Asp Ile
65           70           75           80

Pro Ile Val His Arg Val Ile Lys Val His Glu Lys Asp Asn Gly Asp
          85           90           95

Ile Lys Phe Leu Thr Lys Gly Asp Asn Asn Glu Val Asp Asp Arg Gly
          100          105          110

Leu Tyr Lys Glu Gly Gln Asn Trp Leu Glu Lys Lys Asp Val Val Gly
          115          120          125

Arg Ala Arg Xaa Phe Leu Pro Tyr Val Gly Met Val Thr Ile Ile Met
          130          135          140

Asn Asp Tyr Pro Lys Phe Xaa Tyr Ala Leu Leu Ala Val Met Gly Ala
145          150          155          160

Tyr Val Leu Leu Lys Arg Glu Ser
          165

```

(2) INFORMATION FOR SEQ ID NO:9:

- (i) SEQUENCE CHARACTERISTICS:
 (A) LENGTH: 309 base pairs
 (B) TYPE: nucleic acid
 (C) STRANDEDNESS: double
 (D) TOPOLOGY: linear